

إعداد وتأليف

محمد بشر زينه

الإصدار الأول

إلى أمي وأبي وزوجتي وابنتي وأخي

إلى كل شهيد روى سوريا بدمائه الطاهرة

بسم الله الرحمن الرحيم

مقدمة الكتاب:

بدأت فكرة هذا الكتاب من تدريسي الجانب العملي لطلاب الإحصاء الرياضي في كلية العلوم بجامعة حلب –طلاب السنة الرابعة /مقرر البرامج الإحصائية المتقدمة- مع الدكتور القدير حيان حسن، حيث اقترح علي زملائي وطلابي جمع محاضرات العملي التي كتبتها في كتاب صغير يكون مرجعاً مساعداً لطلاب الإحصاء بجامعة حلب، وحالما بدأت بجمع هذه المحاضرات وترتيبها، بدأت فكرة توسيع الكتاب ليغطي جوانب أخرى هامة في الإحصاء حتى وصل إلى شكله الحالي الذي هو بين يديك.

ينقسم هذا الكتاب إلى ثلاثة عشر فصلاً، حيث تم عرض أساسيات اللغة وبنيتها في أول فصلين، وتم عرض القواعد البرمجية وأساليب التعامل مع البيانات في الفصلين الثالث والرابع، وتم تخصيص الفصل الخامس لتعلم الرسم باستخدام R، والفصل السادس انفرد بمواضيع متعلقة بنظرية الاحتمالات، أما الفصول الستة التالية فكانت تستهدف أهم الأساليب والاختبارات الإحصائية، وفي الفصل الأخير تم التطرق لموضوع التنبؤ باستخدام السلاسل الزمنية.

أتمنى أن أكون قد وفقت في تبسيط أسلوب هذا الكتاب، راجياً المولى عزّ وجلّ أن تتحقق الفائدة المرجوة منه.

حلب 2017/9/22

المؤلف

الفهرس

الفصل الأول
أساسيات لغة R
(Principles of R Language)

مقدمة (Introduction):

لغة البرمجة الإحصائية R هي لغة مفتوحة المصدر (Open Source) ابتكرها روس إيهاكا وروبيرت جنتلمان في جامعة أوكلاند في نيوزيلاندا، ويعود سبب تسميتها بلغة R إلى اسم مبتكريها، وقد صدرت أول نسخة مستقرة للغة R عام 2000.

إن لغة R مبنية على التوابع (functional Language)، أي أنها مزودة بكم هائل من التوابع التي تحل معظم المشاكل التي قد تواجهك، كما أنها لغة برمجة بحد ذاتها تتيح لك فرصة إضافة التوابع التي قد تناسبك أو التي هي من إبداعك الشخصي، وبإمكانك عمل حزمة (package) خاصة بك من التوابع والخوارزميات وإضافتها إلى لغة R رسمياً حتى يستفيد منها غيرك، وهذا هو المقصود بكون R مفتوحة المصدر.

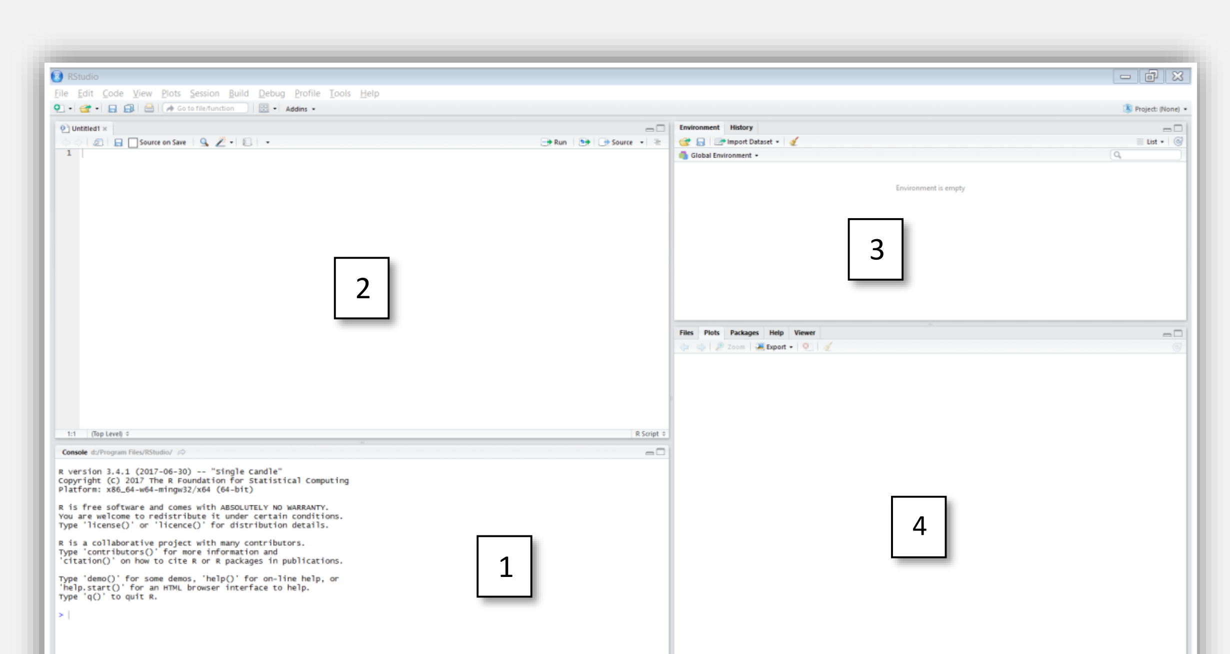
معرفتك وإتقانك للغة البرمجة الإحصائية R أحد أهم الميزات التي يمكن أن تضيفها لسيرتك الذاتية عندما تفكر بالتقدم إلى عمل في مجال الإحصاء أو عندما تفكر في إتمام الدراسات العليا.

يعود سبب انتشار لغة R الواسع والسريع إلى عدة نقاط أهمها:

1- لغة R مجانية.
2- لغة R مفتوحة المصدر (Open Source).
3- لغة R بسيطة وسهلة التعامل والفهم.
4- لغة R تدعم كافة أنظمة التشغيل (Windows, Mac, Linux,…).

بإمكانك تحميل نسخة R الخاصة بك من الرابط: https://cran.r-project.org كما بإمكانك بعدها تنزيل بيئة العمل R-Studio وهي بيئة تطوير كثيرة الاستخدام والشيوع بين مستخدمي R بسبب تبسيطها واختصارها لكتابة الأوامر ولدعمها لميزات كثيرة وذلك من الرابط: https://www.rstudio.com/products/rstudio/download/.

بعد تنصيبك لبرنامج R-Studio وتشغيله ستجد أن بيئة العمل تنقسم إلى أربعة أقسام:



القسم الأول الكونسول Console: وفيه يتم تنفيذ الأوامر، وبإمكانك كتابة الكود أو الأمر الذي تريد تنفيذه ثم الضغط على Enter ليتم التنفيذ، ولست بحاجة لحفظ التعليمات بشكل كامل في R Studio لأنه يمتلك ميزه إكمال العبارات التي يمكنك الاستفادة منها بالضغط على زر Tab، فتنبثق قائمة لكل الأوامر القريبة من الأمر الذي بدأت بكتابته فتختار منها ما تشاء.

القسم الثاني محرر المصدر Source Editor: وفيه يمكن كتابة الأوامر، وتعديلها، وحفظها للاستفادة منها لاحقاً، كما يمكنك تنفيذ السطر الذي تشاء منه بالضغط على Ctrl+Enter وتستطيع تنفيذ أي جزء من الكود بتحديده باستخدام الفأرة ثم الضغط أيضاً على Ctrl+Enter.

القسم الثالث ساحة العمل والحافظة والملفات Workspace, History and files: في ساحة العمل يمكن مشاهدة المتحولات التي تم تعريفها، وفي الحافظة تظهر الأوامر التي تم تنفيذها، كما يمكن إعادة تنفيذ أي تعليمة تريد بمجرد النقر عليها نقرتين متتاليتين، أو نقل التعليمة إلى محرر المصدر بالنقر على زر Shift مع نقرتين متتاليتين على التعليمة،

أما الملفات وهي اختصار لمستعرض الملفات، ففيها يتم عرض الموقع من القرص الصلب والذي يتم العمل فيه، وبإمكانك تغيير الموقع إلى أي مسار تريده.

القسم الرابع الرسوم البيانية والحزم والمساعدة Plots, Packages and Help: يتم عرض جميع الرسوم التي قمت برسمها في **Plots** ويمكنك التنقل بين هذه الرسوم وحفظها، أما **الحزمة**، فهي مجموعة من التوابع المعرفة مسبقاً، ويحتوي R على الكثير من الحزم الجاهزة التي لم تترك أي جانب من الإحصاء إلا ودخلت فيه، وفي هذه اللائحة تستطيع تنزيل الحزم من الانترنت وإجراء التحديثات وغير ذلك، أما لائحة **المساعدة** فتقدم لك المساعدة عن أي أمر تقوم بكتابته في صندوق البحث.

فائدة:
من المواقع المفيدة للبحث عن التعليمات وشرحها موقع www.statmethods.net
محرك البحث الخاص بـ R وتعليماتها www.rseek.org

العمليات الحسابية والمنطقية (Mathematical and Logical Operators):

مثل أي لغة برمجة أخرى، تجري لغة R العمليات الحسابية الأساسية البسيطة، والعمليات المنطقية، والموضحة بالجدول الآتي:

العمليات الحسابية Mathematical Operators		
القوة	** أو ^	2^5=32
الضرب والقسمة	* , /	3*2=6 , 10/2=5
باقي القسمة والقسمة الصحيحة	%% , %/%	7%%3=1 , 7%/%3=2
الجمع والطرح	+ , -	3+1=4 , 3-1=2

وللعمليات السابقة أولوية بالتنفيذ كما تم ذكرها بالجدول السابق على الترتيب، إلا أن الأقواس في العملية الرياضية لها أولوية التنفيذ دوماً.

مثال
6/2*(1+2)
الناتج
9

مثال
6/(2*(1+2))
الناتج
1

سنعرض الآن العمليات المنطقية في R:

<table>
<tr><th colspan="2">العمليات المنطقية Logical Operators</th></tr>
<tr><td>المساواة</td><td>==</td></tr>
<tr><td>عدم المساواة</td><td>!=</td></tr>
<tr><td>أصغر</td><td><</td></tr>
<tr><td>أكبر</td><td>></td></tr>
<tr><td>أصغر أو يساوي</td><td><=</td></tr>
<tr><td>أكبر أو يساوي</td><td>>=</td></tr>
<tr><td>"و" المنطقية</td><td>&</td></tr>
<tr><td>"أو" المنطقية</td><td>|</td></tr>
</table>

مثال

```
> 3==5
[1] FALSE
> 3!=5
[1] TRUE
> 3<5
[1] TRUE
> 3>5
[1] FALSE
> 3<=5
[1] TRUE
> 3>=5
[1] FALSE
> TRUE&FALSE
[1] FALSE
> TRUE|FALSE
[1] TRUE
```

ملاحظة:
إن أي سطر ندخل فيه التعليمات في لغة R يبدأ بـ > أما سطر النتائج فيبدأ بـ [رقم النتيجة] وذلك لأن التعليمة قد تعطي أكثر من نتيجة واحدة.

الكائنات (Objects) وبعض الملحوظات حول R:

إن لغة R كباقي لغات البرمجة تحتوي أنواع متعددة من المتغيرات، والمتغير هو مكان في الذاكرة يمكن تخزين البيانات فيه ويمكن الرجوع له واستخدامه أو تعديله متى شئنا، المتغيرات في R تدعى كائنات (Objects)، وهذه الكائنات يتم تخزينها ضمن ما يسمى ساحة العمل (Workspace).

لمعرفة مسار ساحة العمل الحالية نستخدم التعليمة:

```
getwd()
```

ولتغيير مسار ساحة العمل الحالية نستخدم التعليمة:

```
setwd("new path")
```

بإمكانك معرفة الكائنات التي تم تعريفها حتى اللحظة باستخدام التعليمة:

```
ls()
```

كما يمكنك حذف كائن ما وليكن obj باستخدام التعليمة:

```
rm(obj)
```

لأسماء الكائنات قواعد يجب الالتزام بها وهي:

- يمكن أن يحوي اسم المتغير أي من الأحرف الأبجدية a-b أو A-B أو الأرقام 0-9 أو النقطة (.) أو الخط السفلي (_) ولا يحوي أية رموز أخرى (مثل @%$...).
- يبدأ اسم المتغير حصراً بحرف أو نقطة.
- لو بدأنا اسم المتغير بنقطة فلا يجوز أن يتبع النقطة رقم.

لو أردنا تخزين القيمة 11 في المتغير c فيمكن اتباع أحد الطرائق الأربعة الآتية:

```
c<-11
11->c
c=11
assign('c',11)
```

وللمتغيرات في R أنواع عديدة ومنها المنطقي logical، الصحيح integer، الرقمي numeric، المركب complex، المحرف character، وغيرها.

لمعرفة نوع متغير ما مثل x يمكن استخدام إحدى التعليمتين:

```
typeof(x)
class(x)
```

<u>ملاحظة:</u>
إن R حساسة لحالة الأحرف الكبيرة والصغيرة (Case Sensitive) أي أن a غير A كما أن التعليمة q() غير Q().

<u>ملاحظة:</u>
للمتغير المنطقي أحد القيمتين True أو False واختصاراً T و F.

الأشعة (Vectors):

الأشعة في R هي عبارة عن عدة كائنات لها نفس النوع ومخزنة بترتيب محدد.

يمكن تعريف شعاع x فيه القيم 3,4,5 بالشكل:

```
x<-c(3,4,5)
```

حيث يرمز الحرف c إلى الكلمة concatenate والتي تعني "تسلسل".

كما يمكن معرفة عدد عناصر الشعاع x بالتعليمة:

```
length(x)
```

التابعان seq و rep:

<u>أولاً:</u> التابع seq وله الشكل العالم الآتي:

```
seq(from,to,by)
```

وهو تابع يستخدم لتوليد متتالية من الأرقام من from إلى to وبخطوة by، فلو أردنا مثلاً توليد المتتالية: 1,3,5,7,9,11,13,15 نكتب أحد التعليمتين الآتيتين:

```
seq(1,15,2)
seq(from=1,to=15,by=2)
```

ملاحظة:

يمكن توليد متتالية من العدد **from** إلى العدد **to** بخطوة تساوي واحد بالشكل المختصر الآتي:

```
from:to
```

فلو أردنا مثلاً توليد أعداد من 1 إلى 13 نكتب:

```
1:13
```

ثانياً: التابع rep وله الشكل العام الآتي:

```
rep(x,each)
```

حيث x هو العنصر أو الشعاع المراد تكراره و each هو عدد مرات التكرار، فمثلاً لتوليد العناصر:

1 1 2 2 3 3 4 4 5 5 6 6

نكتب التعليمة:

```
rep(1:6,each=2)
```

أما لتوليد العناصر: 6 5 4 3 2 1 6 5 4 3 2 1 نكتب:

```
rep(1:6,2)
```

ولتوليد العناصر: 11 10 8 11 10 8 11 10 8 فنكتب:

```
rep(c(8,10,11),3)
```

أما لتوليد العناصر 11 11 11 10 10 8 فنكتب:

```
rep(c(8,10,11),1:3)
```

بعض التوابع الرياضية والإحصائية الهامة:

الدالة	الشكل العام	الوظيفة
abs	abs(x)	القيمة المطلقة
log	log(x,base=y)	اللوغاريتم ذو الأساس y لـ x
exp	exp(x)	العدد النيبري مرفوعاً للأس x
sqrt	sqrt(x)	جذر x
factorial	factorial(x)	عاملي x
ceiling	ceiling(x)	تقريب x لأقرب عدد صحيح ليس أكبر من x
floor	floor(x)	تقريب x لأقرب عدد صحيح ليس أصغر من x
trunc	trunc(x)	إرجاع القسم الصحيح فقط من x
round	round(x, digits=n)	تقريب x بدقة n عدداً بعد الفاصلة
cos,sin,tan,acos,cosh,…	cos(x),sin(x),…	النسب المثلثية
min	min(x)	أصغر عدد في شعاع x
max	max(x)	أكبر عدد في شعاع x
range	range(x)	المدى للشعاع x
sum	sum(x)	مجموع عناصر الشعاع x
mean	mean(x)	متوسط عناصر الشعاع x
median	median(x)	وسيط عناصر الشعاع x
var	var(x)	تباين عناصر الشعاع x
sd	sd(x)	الانحراف المعياري لعناصر الشعاع x

الفلترة (Filtering) وبعض التطبيقات على الأشعة:

نقصد بالفلترة الوصول لبيانات الشعاع التي تحقق شرطاً أو عدة شروط، وسنستعرض هذا المفهوم من خلال التطبيق الآتي:

1. املأ الشعاع x بالقيم 7,5,6,7,8,9,11.
2. أوجد حجم الشعاع x.
3. أوجد العنصر الخامس من الشعاع x.
4. ضع الشعاع x في الشعاع y مضيفاً له القيم 4,8,9,10,13.
5. اطبع قيم الشعاع y عدا العنصر رقم 12.
6. اطبع أول 3 قيم من الشعاع y.
7. اطبع القيمة الأولى والخامسة والتاسعة من الشعاع y.
8. استبدل القيمة الأولى من الشعاع y بالقيمة 12.
9. استبدل القيمة الثالثة من الشعاع y بمربعها.
10. استبدل أول 3 قيم من الشعاع y بالقيمة 7.
11. استبدل القيمة السابعة والثامنة والتاسعة من الشعاع y بالقيم 7,8,9.
12. أوجد معكوس الشعاع y.
13. استبدل القيم التي هي أكبر من 8 في الشعاع y بالقيمة 4.
14. اطبع آخر 9 قيم من الشعاع y.
15. أضف للعناصر الزوجية في الشعاع y القيمة 1.
16. استبدل العناصر الفردية التي هي أقل من 7 بالقيمة 2.
17. أوجد كلاً من المجموع والمتوسط والوسيط والانحراف المعياري والانحراف المتوسط للشعاع y.
18. أسند نصف قيم الشعاع y للشعاع x1 والنصف الآخر للشعاع x2.
19. ولد الشعاع rep المكون من العناصر 5 5 5 4 4 3 3 3 3 3 3 3 2 2 2 2 2 1 1 1 بطريقة مختصرة.

الحل:

1	> x<-c(7,5,6,7,8,9,11)
2	> length(x)
3	> x[5]
4	> y<-c(x,4,8,9,10,13)
5	> y[-12]
6	>y[1:3]
7	> y[c(1,5,9)]
8	> y[1]=12
9	> y[3]=y[3]^2
10	> y[1:3]<-7
11	> y[c(7,8,9)]<-c(7,8,9)
12	> rev(y)
13	> y[y>8]<-4
14	> y[(length(y)-8):length(y)]
15	> y[y%%2==0]<-y[y%%2==0]+1
16	> y[y%%2!=0 & y<7]<-2
17	> sum(y) >mean(y) >median(y) >sd(y) > sum(abs(y-mean(y)))/length(y)
18	> x1<-y[1:6] > x2<-y[7:12]
19	> rp<-c(rep(1,3),rep(2,5),rep(3,6),rep(4,2),rep(5,3))

التعليمة subset:

يمكن استخدامها للوصول إلى بيانات الشعاع التي تحقق شرطاً محدداً، ولها الشكل:

```
subset(object, condition)
```

فمثلاً إذا أردنا عرض بيانات الشعاع x التي هي أكبر من 5 والتي هي من مضاعفات العدد 4 نكتب:

```
subset(x,x>5 & x%%4==0)
```

ملاحظة:

يمكن إجراء العمليات الحسابية على شعاعين x,y حيث نقصد بالجمع أو الطرح أو الضرب أو القسمة إجراء كل من هذه العمليات عنصراً لعنصر، أما لإيجاد الجداء الداخلي فنكتب:

x %*% y

المصفوفات (Matrices):

يمكن تعريف مصفوفة عناصرها elements مكونة من nrow سطراً و ncol عموداً بالشكل:

```
matrix(elements,nrow,nclo)
```

فلتعريف المصفوفة الآتية مثلاً:

$$\begin{pmatrix} 3 & 5 & 6 \\ 2 & 1 & 7 \\ 8 & 7 & 2 \\ 7 & 8 & 9 \end{pmatrix}$$

نكتب أحد التعليمات المتكافئة الآتية:

```
matrix(c(3,2,8,7,5,1,7,8,6,7,2,9),4,3)
matrix(c(3,5,6,2,1,7,8,7,2,7,8,9),4,3,byrow=T)
matrix(c(3,2,8,7,5,1,7,8,6,7,2,9),c(4,3))
matrix(c(3,5,6,2,1,7,8,7,2,7,8,9),c(4,3),byrow=T)
```

والعمليات الحسابية الأربعة (الجمع والطرح والضرب والقسمة) تعني إجراء كل من هذه العمليات عنصراً لعنصر, أما لإيجاد الجداء حسب مفهوم جداء مصفوفتين فنكتب:

```
A %*% B
```

بعض التوابع المستخدمة مع المصفوفات:

الدالة	الشكل العام	الوظيفة
solve	solve(A)	مقلوب المصفوفة A
t	t(A)	منقول مصفوفة A
det	det(A)	محدد مصفوفة A

<u>ملاحظة:</u>

يمكن إجراء عمليات الفلترة على المصفوفات كما أجريناها على الأشعة.

التعليمة apply:

للتعليمة apply الشكل العام الآتي:

```
apply(a,row_or_column,statement)
```

يمثل الوسيط الأول المصفوفة المستخدمة، أما الوسيط الثاني فيأخذ إما 1 للدلالة على التعامل مع الأسطر أو 2 للدلالة على التعامل مع الأعمدة، والوسيط الثالث نضع به التعليمة التي نريد تطبيقها على كافة الأسطر أو الأعمدة.

فلو أردنا إيجاد مجموع كافة أسطر المصفوفة A نكتب:

```
apply(A,1,sum)
```

ولو أردنا إيجاد متوسط كافة أعمدة المصفوفة A نكتب:

```
apply(A,2,mean)
```

التعامل مع الأسطر والأعمدة:

سنورد فيما يلي أهم التعليمات التي يمكن استخدامها مع المصفوفات والتعامل مع أسطرها وأعمدتها:

الدالة	الوظيفة
A[r ,]	أخذ السطر رقم r من مصفوفة A
A[, c]	أخذ العمود رقم c من المصفوفة A
rbind(A,X)	إضافة شعاع X كسطر جديد للمصفوفة A
cbind(A,X)	إضافة شعاع X كعمود جديد للمصفوفة A
diag(A)	الحصول على عناصر القطر الرئيسي للمصفوفة A
diag(X)	توليد مصفوفة قطرية قطرها الشعاع X وباقي العناصر أصفار
eigen(A)	إيجاد القيم الذاتية والأشعة الذاتية لمصفوفة مربعة A
eigen(A)$values	إيجاد القيم الذاتية لمصفوفة مربعة A
eigen(A)$vectors	إيجاد الأشعة الذاتية لمصفوفة مربعة A

ملاحظة:

يمكن إضافة أسماء لأسطر وأعمدة المصفوفة x باستخدام التعليمتين colnames و rownames فلو كانت لدينا مصفوفة من ثلاثة أسطر وثلاثة أعمدة أسماء أعمدتها c1,c2,c3 وأسماء أسطرها r1,r2,r2 يمكننا إضافة هذه الأسماء بالشكل:

```
colnames(x) <- c("C1","C2","C3")
rownames(x) <- c("R1","R2","R3")
```

تطبيق:

لتكن لدينا المصفوفة:

$$y = \begin{pmatrix} 3 & 5 & 6 \\ 2 & 1 & 7 \\ 8 & 7 & 2 \\ 7 & 8 & 9 \end{pmatrix}$$

والتي يمكن تعريفها بالشكل:

```
y<-matrix(c(3,2,8,7,5,1,7,8,6,7,2,9),4,3)
```

1. اطبع عناصر السطر الثالث.
2. اطبع عناصر العمود الثاني.
3. استبدل عناصر السطر الأول بالقيم 9 7 4.
4. أنشئ ثلاثة أشعة y3 y2 y1 بالاستعانة بالعمود الأول والثاني والثالث من المصفوفة السابقة على الترتيب.
5. أضف العمود 2 4 3 1 للمصفوفة السابقة.
6. أضف السطر 9 8 7 2 للمصفوفة السابقة.
7. أوجد المتوسط الحسابي للسطر الثاني للمصفوفة y.
8. أوجد المتوسط الحسابي لكل أعمدة المصفوفة y.

	الحـــــــــــــــــــــــــل:
1	y[3,]
2	y[, 2]
3	y[1,]<-c(4,7,9)
4	y1<-y[, 1] y2<-y[, 2] y3<-y[, 3]
5	y<-cbind(y,c(1,3,4,2))
6	y<-rbind(y,c(2,7,8,9))
7	mean(y[2,])
8	apply(y,2,mean)

الفصل الثاني
القوائم وأُطّر البيانات
(Lists and Data frames)

القوائم (Lists):

القائمة (List) هي تجمع لعدد من البيانات مختلفة الأنواع، ولتعريف قائمة اسمها lst1 فيها العناصر a=2.5 و b=True و c=1:3 نكتب التعليمة:

lst1<-list("a"=2.5, "b"=T, "c"=1:3)
الناتج:
$a [1] 2.5 $b [1] TRUE $c [1] 1 2 3

ملاحظة:

يمكن استخدام التعليمة str(lst1) لإعطاء معلومات تفصيلية عن القائمة lst1.

ملاحظة:

يمكن تعريف lst1 بالشكل المكافئ الآتي:

```
x<-list(2.5,T,1:3)
```

الوصول لعناصر القائمة:

لنعرف القائمة الآتية:

```
x<-list("name"="bisher","age"=26,"marks"=c(90,92,94))
```

للوصول للحقل name يمكن استخدام أي تعليمة من التعليمات المتكافئة الآتية:

```
x[1]
x[[1]]
x["name"]
x[["name"]]
x$name
x$n
```

كما يمكن الوصول للعلامة 92 بأحد الأشكال الآتية:

```
x[[3]][2]
x$m[2]
x$marks[2]
```

إطار البيانات (Data frame):

يمكن تعريف إطار البيانات Data frame على أنه جدول يحتوي عدة أسطر وعدة أعمدة حيث يمثل كل عمود نوع محدد من البيانات وكل سطر بيانات فرد محدد كما في المثال الآتي:

Name	Married	Average
Auday	False	75
Omar	True	77
Ahmed	False	80
Ammar	False	81

لإدخال الجدول السابق في Data Frame نكتب:

```
Name<-c("Auday","Omar","Ahmed","Ammar")
Married<-c(F,T,F,F)
Average<-c(75,77,80,81)
tbl<-data.frame(Name,Married,Average)
```

يمكن تعديل الجدول السابق باستخدام محرر إطار البيانات بسهولة وديناميكية أكثر باستخدام التعليمة:

```
fix(tbl)
```

يمكن إيجاد إحصاءات عامة حول Data Frame السابقة بالشكل:

```
summary(tbl)
```

الناتج

```
Name          Married          Average
Ahmed:1       Mode :logical    Min.   :75.00
Ammar:1       FALSE:3          1st Qu.:76.50
Auday:1       TRUE :1          Median :78.50
Omar :1       NA's :0          Mean   :78.25
                               3rd Qu.:80.25
                               Max.   :81.00
```

للوصول للسطر رقم r نكتب:

```
tbl[r, ]
```

للوصول للعمود رقم c نكتب:

```
tbl[ ,c]
```

للوصول للعمود الذي اسمه col نكتب:

```
tbl$col
```

مثلاً:

```
tbl$Name
```

التعليمة View والتعليمة stack:

من التعليمات المفيدة والتي يمكن استخدامها مع Data frame هي التعليمة View والتي تقوم بعرض بيانات الـ Data frame بشكل منسق في نافذة مستقلة، أما التعليمة stack فالهدف منها تجميع بيانات Data frame في عمودين الأول هو البيانات والثاني هو تسميات هذا البيانات، فمثلاً لو كان لدينا:

X1	X2	X3
30	35	37
31	34	38
31	36	38

التعليمة View تجعل عرض البيانات السابقة بالشكل الآتي:

	x1	x2	x3
1	30	35	37
2	31	34	38
3	31	36	38

باستخدام التعليمة stack تصبح الـ Data frame من الشكل:

values	ind
30	X1
31	X1
31	X1
35	X2
34	X2
36	X2
37	X3
38	X3
38	X3

استيراد وتصدير البيانات (Importing and Exporting Data):

سنبدأ أولاً باستيراد البيانات حيث توجد عدة تعليمات لاستيراد البيانات إلى R وذلك باختلاف المصدر الذي نريد استيراد البيانات منه.

-لنفترض أننا نريد استيراد البيانات التي هي في القرص D وذات الاسم Data.

أولاً: الاســـــــــــــــــــتيراد		
البرنامج	الصيغة	التعليمة في R
#	csv	mydata <- read.table("d:/data.csv", header=TRUE,sep=",")
Excel	xlsx	library(xlsx) mydata <- read.xlsx("d:/ data.xlsx", sheetName = "mysheet")
SPSS	por	library(Hmisc) mydata <- spss.get("d:/ data.por", use.value.labels=TRUE)
SAS	xpt	library(Hmisc) mydata <- sasxport.get("d:/data.xpt")
Stata	dta	library(foreign) mydata <- read.dta("d:/data.dta")

الآن ننتقل إلى تصدير البيانات حيث توجد عدة تعليمات لتصدير البيانات من R وذلك باختلاف النمط الذي نريد تصدير البيانات وفقه.

-لنفترض أننا نريد تصدير البيانات المخزنة في الكائن mydata إلى القرص D وباسم Data.

ثانياً: التصــــــــــــــــــــــدير		
البرنامج	الصيغة	التعليمة في R
#	csv	write.table(mydata,"d:/data.csv",sep=",")
Excel	xlsx	library(xlsx) write.xlsx(mydata, "d:/data.xlsx")
SPSS	sps	library(foreign) write.foreign(mydata, "d:/data.txt", "c:/data.sps", package="SPSS")
SAS	sas	library(foreign) write.foreign(mydata, "d:/data.txt", "d:/data.sas", package="SAS")
Stata	dta	library(foreign) write.dta(mydata, "d:/data.dta")

إن تصدير البيانات لـ SPSS أو SAS يقتضي تصدير البيانات كنسخة txt أولاً ثم كتابة Syntax للبرنامج المستهدف حتى تتم قراءة البيانات من ملف الـ txt.

الفصل الثالث

العبارات الشرطية والعبارات التكرارية والتوابع

(Conditional Statements, Loops and Functions)

العبارة الشرطية if:

تستخدم العبارة الشرطية if عندما نرغب بتنفيذ تعليمة محددة أو عدة تعليمات statements عندما يتحقق شرط أو عدة شروط conditions، ولها الشكل العام الآتي:

```
if(conditions)
{
statements
}
```

ولها شكل أعم، وهو تنفيذ كتلة تعليمات statements1 في حال تحقق الشروط conditions1 وتنفيذ كتلة التعليمات statements2 في حال عدم تحققها:

```
if(conditions)
{
statements1
} else
{
statements2
}
```

كما يمكن أيضاً للتعليمة if أن تأخذ شكلاً أعم، فتختبر عدة شروط ويكون تحقق كل شرط من الشروط مقروناً بتعليمات خاصة به وذلك بالشكل الآتي:

```
if(conditions1)
{
statements1
} else if(condition2)
{
statements2
} else if (condition3)
...
```

مثال:

برنامج يقوم بطباعة Negative إذا كان العدد المدخل **سالباً**، ويطبع Positive إذا كان العدد المدخل موجباً، ويطبع صفر فيما عدا ذلك:

```
> x<-0
> if (x<0)
+ {
+ print("Negative")
+ } else if(x>0)
+ {
+ print("Positive")
+ } else print("Zero")
```

ملاحظة:
الرمز + في بداية الأسطر من الثالث إلى الأخير تعني أن التعليمة لم تنته، وهنا يجب أن نشير إلى أنه علينا ألا نكتب else إلا بجوار القوس ولا نفردها بسطر وحدها وإلا اعتبرت R أن الكود انتهى.

التابع ifelse:

للتابع ifelse الشكل العام الآتي:

```
ifelse(condition,x,y)
```

وهو تابع يفحص العبارة المنطقية condition فإذا كانت نتيجتها TURE ينفذ التعليمة X وإذا كانت نتيجتها FALSE ينفذ التعليمة Y.

مثال:
a <-c(5,7,2,9) ifelse(a %% 2 == 0,"even","odd")
الناتج:
"odd" "odd" "even" "odd"

العبارة Switch:

للتعليمة switch الشكل العام الآتي:

```
switch(statement, list)
```

حيث يتم إرجاع قيمة من القائمة list بالاعتماد على قيمة statement، فمثلاً:

التعليمة
switch(2,"red","green","blue")
الناتج
green""
التعليمة
switch("color", "color" = "red", "shape" = "square", "length" = 5)
الناتج
red""

حلقة التكرار for:

لحلقة التكرار for الشكل العام الآتي:

```
for(i in start:end)
{
Statements
}
```

حيث:

i: العداد.

start: نقطة البداية.

end: نقطة النهاية.

statements: التعليمات المراد تكرارها.

تقوم تعليمة for بتكرار التعليمات Statements بعدد من المرات يساوي end-start+1.

مثال: لتكن المصفوفة:

$$x = \begin{pmatrix} 1 & 5 & 6 \\ 2 & 7 & 6 \\ 4 & 3 & 7 \end{pmatrix}$$

والمطلوب:

1. إدخال المصفوفة.
2. طباعة عناصر القطر الرئيسي.
3. طباعة عناصر القطر الثانوي.
4. طباعة العناصر التي هي فوق القطر الرئيسي.
5. طباعة العناصر التي هي فوق القطر الرئيسي مع القطر الرئيسي.
6. طباعة العناصر التي هي تحت القطر الرئيسي.
7. طباعة العناصر التي هي فوق القطر الثانوي.
8. طباعة العناصر التي هي تحت القطر الثانوي مع القطر الثانوي.

	الحـــــــــــــــــــــــل
1.	> x<-matrix(c(1,2,4,5,7,3,6,6,7),3)
2.	> for (i in 1:3) + print(x[i,i]) أو diag(x)
3.	> for (i in 1:3) + print(x[i,3-i+1])
4.	> for (i in 1:2) + print(x[i,(i+1):3])
5.	> for (i in 1:3) + print(x[i,i:3])
6.	> for (i in 2:3) + print(x[i,1:(i-1)])
7.	> for (i in 1:2) + print(x[i,1:(3-i)])
8.	> for (i in 1:3) + print(x[i,(3-i+1):3])

حلقة التكرار while:

لحلقة التكرار while الشكل العام الآتي:

```
while(conditions)
{
statements
}
```

حيث:

conditions: شروط تنفيذ الحلقة.

statements: التعليمات التي تتنفذ في حال تحقق شروط تنفيذ الحلقة.

والبرنامج الآتي يطبع الأعداد من ١ إلى 9:

```
> i<-1
> while (i<10)
+ {
+ print(i)
+ i=i+1
+ }
```

التعليمتان break و next:

تستخدم التعليمة break ضمن حلقة تكرار لإيقافها عند تحقق شرط محدد، والبرنامج الآتي يطبع الأعداد من 1 إلى 10 ويتوقف عن الطباعة عند أول عدد زوجي من مضاعفات العدد 3:

```
> for (i in 1:10)
+ {
+ if (i%%2==0 & i%%3==0) break
+ print(i)
+ }
```

بينما تستخدم التعليمة next لتجاهل تنفيذ التعليمات التي تليها في حلقة ما دون الخروج من الحلقة، والبرنامج الآتي يقوم بطباعة الأعداد الواقعة بين ١ و 25 والتي تقبل القسمة على 3 فقط:

```
> for (i in 1:25)
+ {
+ if (i%%3!=0) next
+ print(i)
+ }
```

حلقة repeat:

وهي حلقة يجب استخدام التعليمة break معها حصراً لإيقاف تنفيذها، ولها الشكل العام الآتي:

```
repeat
{
statements
}
```

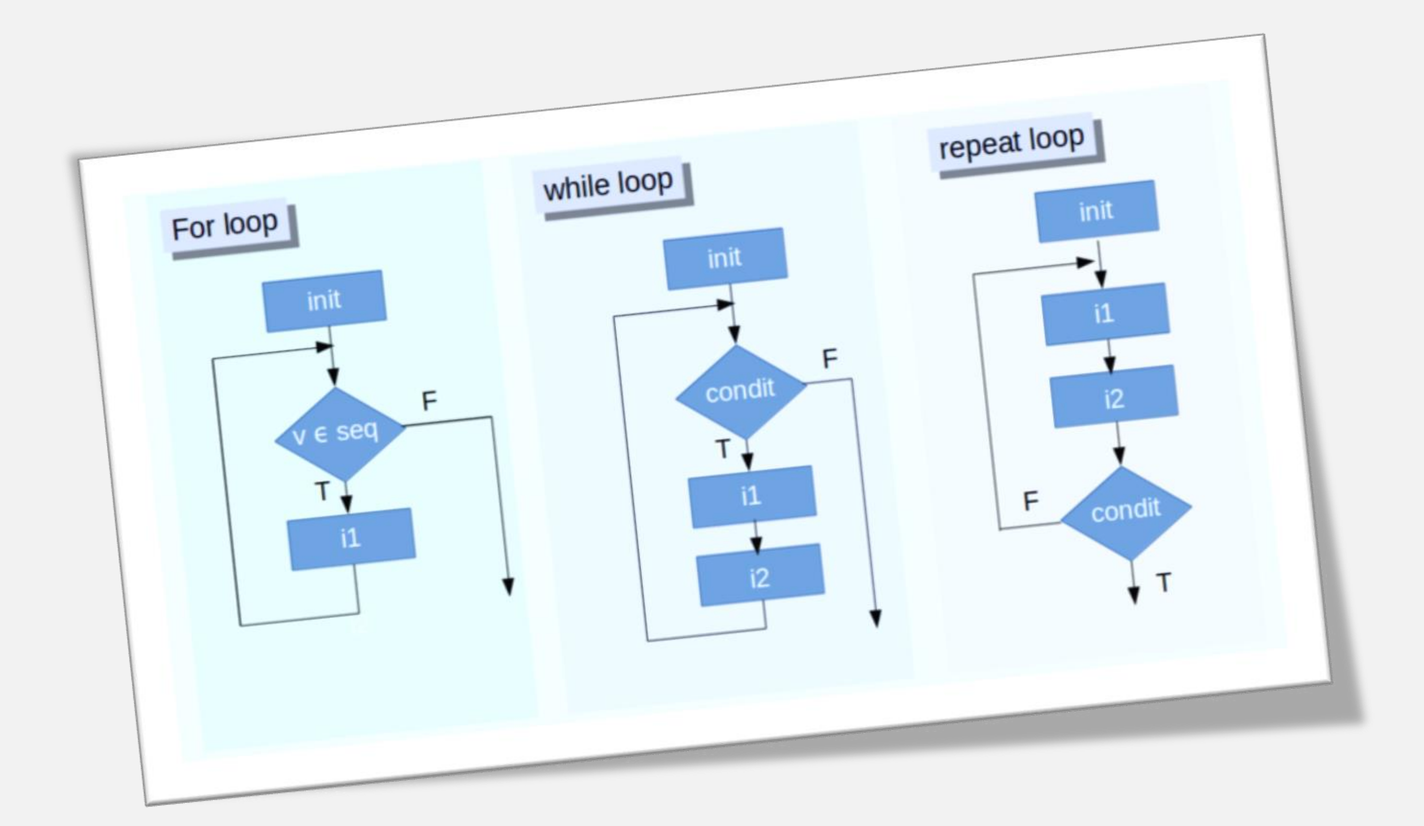
البرنامج الآتي يطبع الأعداد من 1 إلى 10:

```
> i<-1
> repeat
+ {
+ print(i)
+ if (i==10) break
+ i<-i+1
+ }
```

خلاصة الحلقات التكرارية الثلاثة:

- الحلقة for تكرر التعليمات Statements طالما أن العداد ضمن المجال seq المحدد له.
- الحلقة while تكرر التعليمات Statements طالما أن شرط التنفيذ Condition محقق.
- الحلقة repeat تتنفذ مرة واحدة على الأقل إلى أن يتحقق شرط الخروج Condition.

وهذا ما يوضحه الشكل الآتي:



التوابع (functions):

تستخدم التوابع لتقسيم الكود إلى عدة أجزاء بسيطة مما يسهم في تبسيط الكود وفهمه أكثر، ويعرف التابع بالشكل الآتي:

```
func_name <-function(arguments)
{
statements
}
```

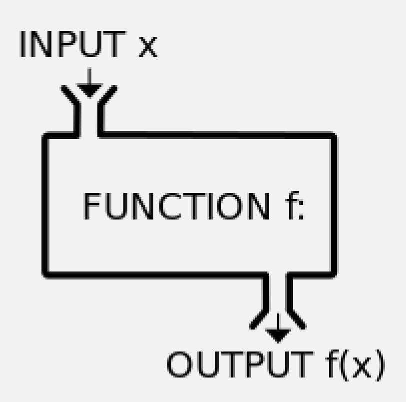
حيث:

func_name: اسم التابع.

arguments: وسطاء التابع.

statements: تعليمات.

الشكل التوضيحي الآتي يساعد على فهم أبسط للتوابع:



تمثل INPUT وسطاء التابع وهي المدخلات التي سنعطيها للتابع FUNCTION f ليجري عليها تعليمات Statements محددة ثم يعطينا المخرجات OUTPUT وهي تحمل اسمه f(x).

مثال:

تابع يستفيد من تابعين لحساب معامل الالتواء لعينة X بالاعتماد على القانون:

$$Skew = \frac{\mu_3^2}{\mu_2^3}$$

حيث:

$$\mu_2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n}$$

$$\mu_3 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^3}{n}$$

```
> mu2<-function(x)
+ {
+ sum((x-mean(x))^2)/length(x)
+ }
> mu3<-function(x)
+ {
+ sum((x-mean(x))^3)/length(x)
+ }
> skew<-function(x)
+ {
+ mu3(x)^2/mu2(x)^3
+ }
```

ويفضل استخدام التعليمة return مع التوابع التي تعيد قيمة محددة كما في المثال الآتي حيث سنعرف تابعاً اسمه check يعيد إما Positive أو Negative أو Zero حسب العدد المدخل:

```
> check <- function(x) {
+   if (x>0) {
+     return("Positive")
+   }
+   else if (x<0) {
+     return("Negative")
+   }
+   else {
+     return("Zero")
+   }
+ }
```

ملاحظة:
التابع هو بنية تعيد قيمة وحيدة فقط (Object)، وإذا أردنا إعادة أكثر من قيمة يمكن جعل التابع يعيد قائمة (List) مثلاً أو أي نوع آخر.

مثال

```
> multi_return <- function() {
+    my_list <- list("color" = "red", "size" = 20, "shape" = "round")
+    return(my_list)
+ }
> a <- multi_return()
> a
```

الناتج

```
$color
[1] "red"

$size
[1] 20

$shape
[1] "round"
```

الفصل الرابع
الأصناف
(Classes)

إن R تدعم البرمجة كائنية التوجه، بل إن كل شيء في R هو عبارة عن كائن، نستطيع وصف الـ Class على أنها بنية معطيات تمتلك خواصاً وتوابع خاصة (طرائق) (methods) بها لتحدد ميزات جميع الكائنات التي سنشتقها منها.

نستطيع تخيل الـ Class على أنها مخطط هيكلي لمنزل وهذا المخطط يحوي جميع تفاصيل المنزل من أرض وسقف وأبواب وشبابيك و ....، وبالاعتماد على هذه التفاصيل نستطيع بناء المنزل.

المنزل الذي نبنيه بالاعتماد على هذه الـ Class هو الـ Object المشتق منها، وبالطبع نستطيع بناء أكثر من منزل بالاعتماد على نفس المخطط وكذلك الأمر بالنسبة لاشتقاق الكائنات المتعددة من الـ Class الأساسية.

الفرق في R عن باقي لغات البرمجة أنها تدعم ثلاثة أنواع من الأصناف وهي S3 و S4 و الأصناف المرجعية.

الصنف S3 هو أبسط أنواع الأصناف وتعريفه بسيط جداً وليس له شكل تصريح رسمي إنما يتم باستخدام كلمة class فقط مما جعل استخدامه شائعاً في R، ويمكن تعريف صنف اسمه Student يملك المكونات name=John وage=21 وaverage=88 بالشكل الآتي:

```
s <- list(name = "John", age = 21, GPA = 3.5)
class(s) <- "student"
```

الصنف S4 هو تحسين على الصنف S3 وله شكل تصريح رسمي يسهل عملية إنشاء كائنات من نفس الصنف تكون أقل أو أكثر شبهاً به، ويتم تعريف مكونات الصنف باستخدام التعليمة setClass() ويتم إنشاء كائنات منه باستخدام التعليمة new()، الصنف السابق يصبح بالشكل:

```
setClass("student", slots=list(name="character", age="numeric",
average="numeric"))
```

الأصناف المرجعية تمت إضافتها حديثاً إلى R وهي أكثر شبهاً بالأصناف في لغات البرمجة كائنية التوجه مقارنة مع الصنفين السابقين، ويمكن تعريفها على أنها أصناف S4 ببيئة خاصة مضافة لها، وتعرف بالشكل الآتي:

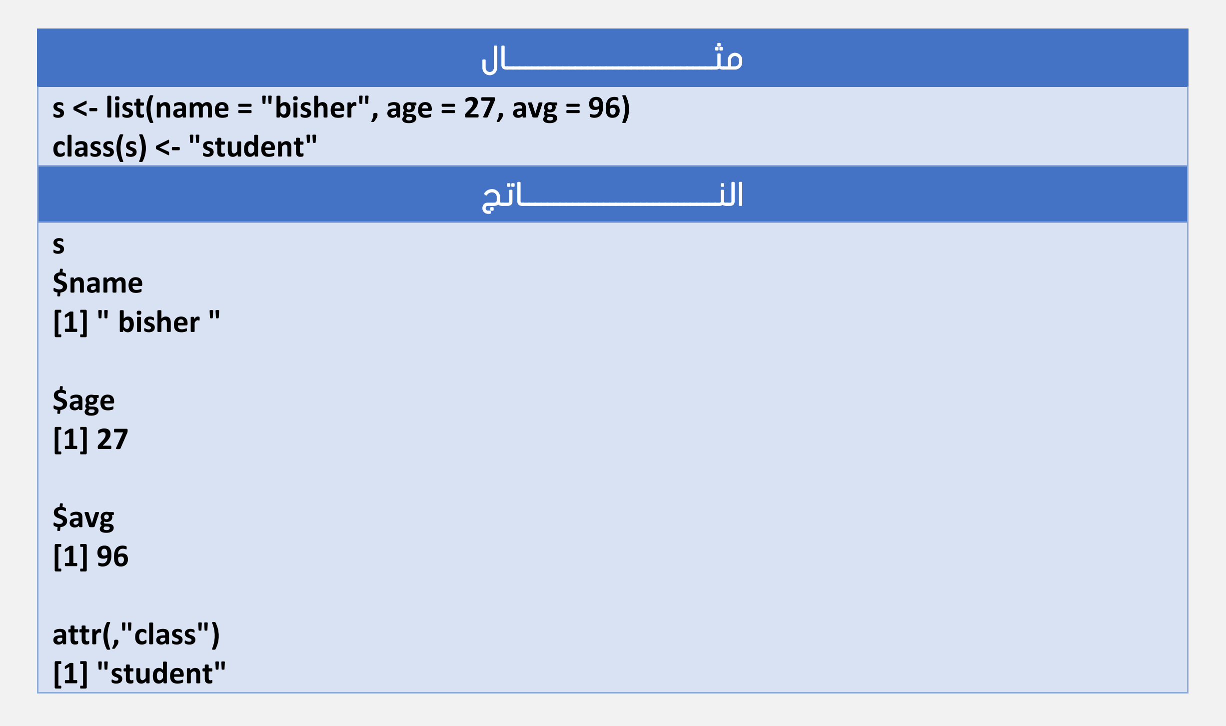
```
setRefClass("student")
```

وسنشرح فيما يلي أنواع الأصناف السابقة بالتفصيل:

الصنف S3:

كما سبق وقلنا إن النمط S3 هو أكثر أنماط الأصناف انتشاراً في R ومعظم الأصناف مسبقة التعريف في لغة R هي من النوع S3 ويعود ذلك لسهولة استخدام هذا النمط.

ليس لهذا النمط صيغة رسمية لتعريفه معرفة مسبقاً في لغة R، لكن نستطيع أن نقول إن القائمة (List) المسندة لاسم Class محددة هي كائن من النوع S3، وعندها تكون مركبات القائمة هي أعضاء الصنف الجديد المعرف.



مثـــــــــــــــال

```
s <- list(name = "bisher", age = 27, avg = 96)
class(s) <- "student"
```

النـــــــــــــــاتج

```
s
$name
[1] " bisher "

$age
[1] 27

$avg
[1] 96

attr(,"class")
[1] "student"
```

استخدام المشيدات Constructors لإنشاء الكائنات:

من المفيد استخدام توابع تحمل نفس اسم الـ Class لإنشاء الكائنات، والذي سيعطينا انتظاماً لإنشاء الكائنات وجعلها تبدو متشابهة.

مثـــــــــــال
student <- function(n,a,av) { if(av>100 || av<0) stop("avg must be between 0 and 100") value <- list(name = n, age = a, avg = av) attr(value, "class") <- "student" value }

استخدمنا في المثال السابق التابع ()attr لجعل value يتبع للصنف student وهي تكافئ تعليمة class(value)<-student.

فيما يلي النواتج المختلفة باختلاف المدخلات للمثال السابق:

التعليمة
s <- student("Paul", 26, 77) s
الناتج
$name [1] "Paul" $age [1] 26 $avg [1] 77 attr(,"class") student"" [1]
التعليمة
s<-student("Paul",25,105)
الناتج
Error in student("Paul", 25, 105) : avg must be between 0 and 100

الطرائق والتوابع العامة (Methods and Generic Functions):

لاحظنا في الأمثلة السابقة أننا وبمجرد ذكر اسم الكائن تتم طباعة كافة محتوياته، وذلك يتم باستدعاء تلقائي للتابع ()print.

يمكننا أيضاً استخدام التابع ()print مع الأشعة vectors والمصفوفات matrices وأطر البيانات data frames و...، **والسؤال الآن:** كيف يعرف التابع ()print كيفية طباعة هذه الكائنات المختلفة وغير المتشابهة؟؟!

الجواب: إن ()print تابع عام generic function، أي أنه تابع يمتلك العديد من الطرائق ويمكنك معرفتها باستخدام التعليمة:

```
methods(print)
```

ومن الطرائق التي يمتلكها التابع ()print نجد مثلاً print.data.frame وبالتالي عند استدعاء التابع ()print مع data frame يتم إرسال التعليمة إلى الطريقة ()print.data.frame.

أي أن الشكل العام للطرائق هو:

```
generic_name.class_name()
```

أي أن استدعاءنا للكائن من الصنف student ينبغي أن يستدعي طريقة من الشكل ()print.student، لكن هذه الطريقة ليست موجودة! **إذاً ما هي الطريقة التي يستدعيها الكائن من الصنف student؟؟!**

إن الطريقة التي يستدعيها الكائن من الصنف student هي ()print.default والتي نسميها الطريقة الاحتياطية أو الافتراضية والتي يتم استدعاؤها عند عدم وجود مطابقة مع المطلوب استدعاؤه.

السؤال الآن.. كيف بإمكاننا كتابة طرائقنا الخاصة؟

الكود الآتي يقوم بتضمين طريقة جديدة باسم student:

```
print.student <- function(obj) {
 cat(obj$name, "\n")
 cat(obj$age, "years old\n")
 cat("Average:", obj$avg, "\n")
}
```

وعند طباعة محتويات الكائن من النوع student سوف تتنفذ هذه الطريقة مباشرة.

في الأصناف من النمط S3 لا تنتمي الطرائق إلى كائن محدد أو صنف محدد إنما تنتمي إلى التوابع العامة generic functions.

كيف يمكننا كتابة تابع عام generic function خاص بنا؟

يجب أن نعلم أن التابع العام هو تابع يملك استدعاءً التابع ()UseMethod مع تمرير اسم التابع له، والذي يمثل التابع المُرسِل الذي سيعالج جميع التفاصيل.

سنعرف الآن تابعاً عاماً اسمه grade وذلك بالشكل الآتي:

```
grade <- function(obj) {
UseMethod("grade")
}
```

ليس للتابع العام أي فائدة بدون ربطه بطرائق خاصة به، ولذلك سنعرف الطريقة الافتراضية أولاً:

```
grade.default <- function(obj) {
cat("This is a generic function\n")
}
```

ثم سنعرف طريقة خاصة بالصنف student:

```
grade.student <- function(obj) {
 cat("Your average is", obj$avg, "\n")
}
```

وبناء عليه يكون الناتج:

```
> grade(s)
Your average is 77
```

الصنف S4:

ما يميز الصنف S4 عن الصنف S3 هو أن للأصناف من النوع S4 نموذج محدد لتعريفها، وذلك يتم بالاعتماد على التابع ()setClass، ونسمي المتحولات الأعضاء للصنف باسم الشرائح (Slots)، ولتعريف الصنف student الذي له الشرائح name ageg avgg نكتب:

```
setClass("student", slots=list(name="character", age="numeric", avg="numeric"))
```

أما إنشاء كائنات من الصنف S4 فيتم باستخدام التابع ()new وفق الشكل الآتي:

```
s <- new("student",name="John", age=21, avg=88)
```

وعند طباعة الكائن s سيكون الناتج:

```
An object of class "student"
Slot "name":
[1] "John"

Slot "age":
[1] 21

Slot "avg":
[1] 88
```

ملاحظة:

يمكننا التأكد فيما إذا كان كائن ما من النمط S4 باستخدام التابع المنطقي ()isS4 فيعيد التابع TRUE إذا كان الكائن من صنف S4 ويعيد FALSE فيما عدا ذلك.

إن التابع ()setClass يعيد تابعاً مولداً له عمل يشابه كثيراً عمل المشيدات ويستخدم لإنشاء كائنات جديدة، فلو كتبنا:

```
student <- setClass("student", slots=list(name="character", age="numeric",
avg="numeric"))
```

لأصبح لدينا تابعاً مولداً اسمه ()student نستطيع استخدامه في إنشاء كائنات جديدة من الصنف student كما في المثال الآتي:

التعليمة
student(name="John", age=21, avg=85)
الناتج
An object of class "student" Slot "name": [1] "John" Slot "age": [1] 21 Slot "avg": [1] 85

يمكننا الوصول إلى شرائح الكائن باستخدام المعامل $ فمثلاً:

التعليمة
s@name
الناتج
[1] "John"

كذلك بإمكاننا التعديل على الشرائح بنفس الأسلوب كما يلي:

التعليمة
s@avg <- 77 s
الناتج
An object of class "student" Slot "name": [1] "John" Slot "age": [1] 21 Slot "avg": [1] 77

كما يمكن إجراء التعديل باستخدام التابع ()slot كما يلي:

التعليمة
slot(s,"name") <- "Paul" s
الناتج
An object of class "student" Slot "name": [1] "Paul" Slot "age": [1] 21 Slot "avg": [1] 77

الطرائق والتوابع العامة (Methods and Generic functions):

لاحظنا في المثال السابق أن مجرد كتابة اسم الكائن s أدت إلى طباعة جميع محتوياته، وذلك يتم بفضل التابع العام ()show والذي يشابه التابع ()print في الصنف من النمط S3.

وبنفس التسلسل الذي اتبعناه في الصنف من النمط S3 سنعرض الآن طريقة تعريف طرائقنا الخاصة للأصناف من النمط S4:

لنعرف الآن الطريقة العامة ()show على سبيل المثال:

```
setMethod("show",
        "student",
        function(object) {
          cat(object@name, "\n")
          cat(object@age, "years old\n")
          cat("avg:", object@avg, "\n")
        }
)
```

والمثال الآتي يوضح الاختلاف بعد التعريف السابق:

التعليمة
s <- new("student",name="John", age=21, avg=90) s
الناتج
John 21 years old avg: 90

وبهذا الأسلوب نستطيع كتابة طرائق الصنف S4 للتوابع العامة.

الأصناف المرجعية (Reference Classes):

لاحظنا في كل من الأصناف من النمط S3 والأصناف من النمط S4 أن الطرائق تنتمي إلى التوابع العامة (Generic functions)، لكن في الأصناف المرجعية تنتمي الطرائق إلى الصنف بحد ذاته مما يجعلها أقرب إلى البرمجة كائنية التوجه كما في لغات البرمجة C++, Java,...

يمكننا تعريف صنف مرجعي باستخدام التابع setRefClass() ونسمي الأعضاء التابعة للأصناف المرجعية بالحقول (fields) ولتعريف صنف مرجعي اسمه student فيه الحقول name,age,avg نكتب:

```
setRefClass("student", fields = list(name = "character", age = "numeric", avg = "numeric"))
```

إن التابع setRefClass() يعيد تابعاً مولداً يستخدم لتوليد كائنات من الصنف المعرف:

التعليمة
student <- setRefClass("student", fields = list(name = "character", age = "numeric", avg = "numeric")) s<-student(name = "John", age = 21, avg = 63) s
الناتج
Reference class object of class "student" Field "name": [1] "John" Field "age": [1] 21 Field "avg": [1] 63

يمكننا الوصول إلى حقول الكائن باستخدام المعامل $ فمثلاً:

التعليمة
s@name
الناتج
[1] "John"

كذلك بإمكاننا التعديل على الشرائح بنفس الأسلوب كما يلي:

التعليمة
s@name <- "Rami" s
الناتج
Reference class object of class "student" Field "name": [1] "Rami" Field "age": [1] 21 Field "avg": [1] 63

ملاحظة هامة:

عند إسناد كائن إلى متحول جديد فإن عملية الإسناد تتم بالقيمة (by value) فمثلاً:

التعليمة
a <- list("x" = 1, "y" = 2) b <- a b$y = 3 a b
الناتج
$x [1] 1 $y [1] 2 $x [1] 1 $y [1] 3k

أي أن التغيير طرأ على b فقط ولم يطرأ أي تغيير على قيم a وهذا يعرف بالمرور بالقيمة، لكن هذا الكلام لا ينطبق على الكائنات المرجعية، حيث توجد نسخة وحيدة من الكائن،

وكل المتحولات الأخرى تشير لنفس الكائن، ومن هنا أتت تسمية الأصناف المرجعية بهذا الاسم، وسنوضح هذا بالمثال الآتي:

التعليمة
a <- student(name = "John", age = 21, avg = 75) b <- a b$name <- "Paul" a b
الناتج
Reference class object of class "student" Field "name": [1] "Paul" Field "age": [1] 21 Field "avg": [1] 75 Reference class object of class "student" Field "name": [1] "Paul" Field "age": [1] 21 Field "avg": [1] 75

ونلاحظ أن تعديلنا على b أدى إلى تغيير قيمة a وهذا ما يعرف بالمرور بالمرجع.

وإذا أردنا عمل نسخة مستقلة بحيث أن التعديل عليها لا يؤثر على الكائن الأساسي نستخدم التابع ()copy كما في المثال الآتي:

التعليمة
a <- student(name = "John", age = 21, avg = 75) b <- a$copy() b$name <- "Paul" a b

الناتج

```
Reference class object of class "student"
Field "name":
[1] "John"
Field "age":
[1] 21
Field "avg":
[1] 75

Reference class object of class "student"
Field "name":
[1] "Paul"
Field "age":
[1] 21
Field "avg":
[1] 75
```

ونلاحظ من المثال السابق أن التغيير جرى على b فقط.

الطرائق المرجعية:

سبق وقلنا إن الطرائق في الأصناف المرجعية تنتمي إلى الصنف وليس إلى التوابع العامة، وإن جميع الأصناف المرجعية تمتلك طرائق معرفة تلقائياً تتم وراثتها من الصنف الرئيسي envRefClass (سنشرح معنى "وراثة" في فقرة لاحقة).

ومن الطرائق المعرفة تلقائياً نذكر ...,field(),show()، copy()، وبإمكاننا بالطبع إضافة طرائقنا الخاصة وذلك عند تعريف الصنف بتمرير قائمة من التوابع للخاصية methods للـ setRefClass() كما في المثال الآتي:

مثـــــــــــــــال

```
student <- setRefClass("student",
  fields = list(name = "character", age = "numeric", avg = "numeric"),
  methods = list(
    inc_age = function(x) {
      age <<- age + x
    },
    dec_age = function(x) {
      age <<- age - x
    }
  )
)
```

في المثال السابق قمنا بتعريف طريقتين سميناهما ()inc_age و ()dec_age واللتان ستعدلان على الحقل age.

<u>لاحظ</u> أننا استخدمنا الإسناد غير المحلي ->> وذلك لأن age ليس خاصاً بالطريقة المعرفة فقط، حيث أن استخدام الإسناد المحلي -> سينشئ متحولاً محلياً خاصاً بالطريقة اسمه age.

ولو قمنا بتنفيذ الكود السابق لكانت لدينا النتائج الآتية:

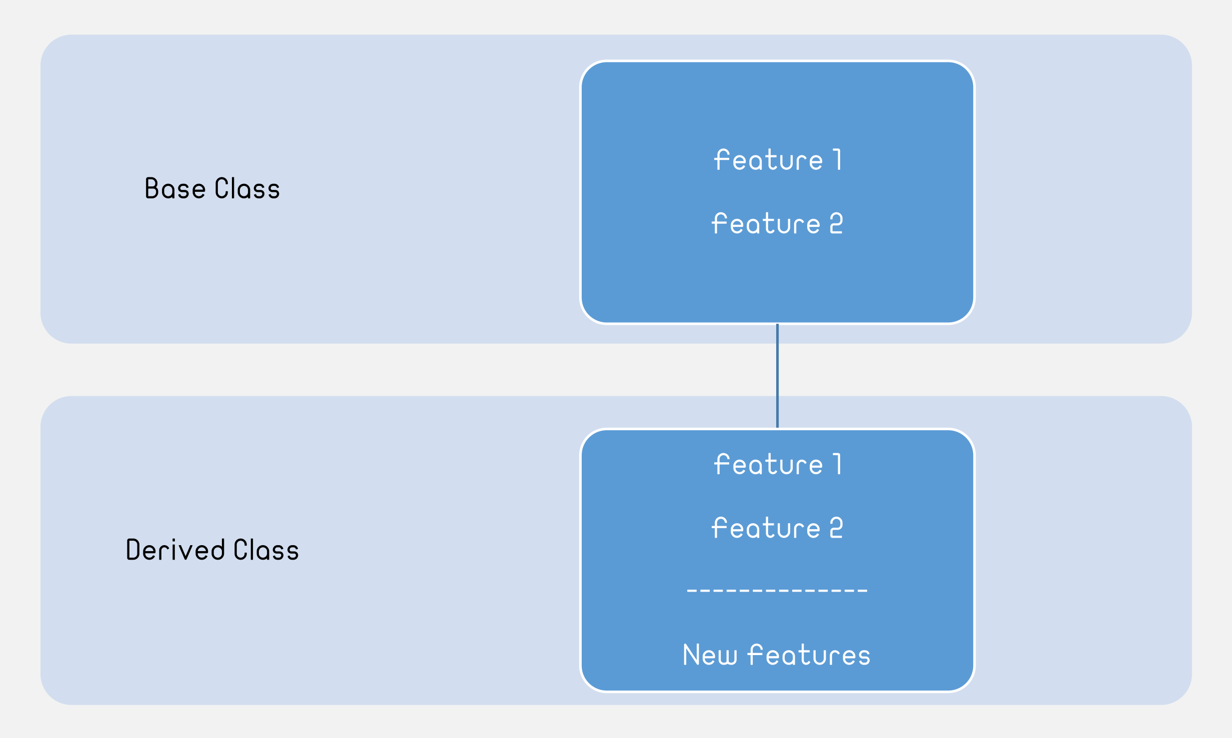
التعليمة
s <- student(name = "John", age = 21, avg = 75) s$inc_age(5) s$age s$dec_age(10) s$age
الناتج
26 16

مقارنة بين الأصناف الثلاثة:

S3 Class	S4 Class	Reference Class
ليس لها صيغة رسمية للتعريف	تعرف باستخدام()setClass	تعرف باستخدام()setRefClass
يمكن اشتقاق الكائنات بتعديل محتويات الصنف الأساسي	يتم اشتقاق كائنات باستخدام التعليمة()new	يتم اشتقاق الكائنات باستخدام التوابع المولدة
الوصول للمعطيات يكون باستخدام المعامل $	الوصول للمعطيات يكون باستخدام المعامل @	الوصول للمعطيات يكون باستخدام المعامل $
الطرائق تنتمي إلى التوابع العامة	الطرائق تنتمي إلى التوابع العامة	الطرائق تنتمي إلى الصنف
تمتاز طرائقها بخاصية المرور بالقيمة	تمتاز طرائقها بخاصية المرور بالقيمة	تمتاز طرائها بخاصية المرور بالمرجع

الوراثة (Inheritance):

من أهم الخواص التي تمتاز بها البرمجة كائنية التوجه هي الوراثة، والتي تعني أننا نستطيع اشتقاق صنف جديد من صنف **أساسي** (Base Class) وإضافة ميزات جديدة للصنف المشتق (Derived Class) مع الاحتفاظ بميزات الصنف الأ**ساسي** دون الحاجة لإعادة كتابتها مرة أخرى.



أي أن الوراثة تشكل تسلسلاً هرمياً للأصناف مثل شجرة العائلة بحيث يكون الصنف الأ**ساسي** في رأس الهرم والأصناف المشتقة تحته بالتسلسل.

سنناقش الآن الوراثة في الأصناف الثلاثة التي تعرفنا عليها **س**ابقاً و**س**نعتمد على الصنف student الذي عرفناه **س**ابقاً:

الوراثة من الصنف S3:

كنا قد عرفنا الصنف student بالشكل:

```
s <- list(name = "bisher", age = 27, avg = 96)
class(s) <- "student"
```

وسنضيف تابعاً ينشئ كائناً من هذا الصنف بالشكل:

```
function(n,a,g) {
  value <- list(name=n, age=a, avg=av)
  attr(value, "class") <- "student"
  value
}
```

ولنعرف الآن طريقة للتابع العام ()print كما يلي:

```
print.student <- function(obj) {
  cat(obj$name, "\n")
  cat(obj$age, "years old\n")
  cat("avg:", obj$avg, "\n")
}
```

سنقوم بتوريث الصنف السابق إلى صنف جديد نسميه InheritedStudent ونشكل منه كائناً وذلك وفق الشكل العام الآتي:

```
class(obj)<-c(child,parent)
```

حيث obj: هو الكائن المشكل من الصنف الموروث.

child: هو الصنف الموروث.

parent: هو الصنف الأساسي الذي سنرث منه.

وبالتالي في مثالنا سنكتب:

التعليمة
s <- list(name="John", age=21, avg=95, country="France") ","student") class(s) <- c("InheritedStudent s
الناتج
John 21 years old avg: 95

ونلاحظ مما سبق أنه تم استدعاء الطريقة الموروثة print.student() كوننا لم نعرف طريقة جديدة للتابع العام print()، ولنعرف الآن طريقة جديدة له باسم print.InheritedStudent() بالشكل الآتي:

التعليمة
print.InheritedStudent <- function(obj) {   cat(obj$name, "is from", obj$country, "\n") } s
الناتج
John is from France

يمكننا معرفة فيما إذا كان الكائن هو من صنف موروث باستخدام أحد التابعين المنطقيين inherits() أو is() كما في الشكل الآتي:

التعليمة
inherits(s,"student")
الناتج
TRUE
التعليمة
is(s,"student")
الناتج
TRUE

الوراثة من الصنف S4:

سنعرف أولاً صنفاً اسمه student مع طريقة للتابع العام show():

```
setClass("student",
  slots=list(name="character", age="numeric", avg="numeric")
)
setMethod("show", "student",
  function(object) {
    cat(object@name, "\n")
    cat(object@age, "years old\n")
    cat("avg:", object@avg, "\n")
  }
)
```

الوراثة في الأصناف S4 تتم عند تعريف الصنف الموروث باستخدام التعبير contains كما يلي:

```
setClass("InheritedStudent",
  slots=list(country="character"),
  contains="student"
)
```

وبالتالي عند تعريف كائن جديد من الصنف الموروث سيكون لدينا:

التعليمة
s <- new("InheritedStudent", name="John", age=21, avg=95, country="France") show(s)
الناتج
John 21 years old avg: 95

ونلاحظ أن الطريقة المعرفة للصنف student تم استدعاؤها عند تنفيذ التعليمة show(s)، وكما سبق في الأصناف S3 نستطيع هنا أيضاً تعريف طرائق للأصناف الموروثة ليتم تنفيذها بدلاً من الطرائق الموروثة من الصنف الأساسي.

الوراثة من الأصناف المرجعية:

الوراثة من الأصناف المرجعية تشبه كثيراً الوراثة من الأصناف S4 حيث علينا فقط إضافة التعليمة contains في التصريح عن الصنف الموروث.

سنعرف أولاً صنفاً اسمه student ونعرف فيه طريقتين أسماؤها inc_age و dec_age:

```
student <- setRefClass("student",
  fields=list(name="character", age="numeric", avg="numeric"),
  methods=list(
    inc_age = function(x) {
      age <<- age + x
    },
    dec_age = function(x) {
      age <<- age - x
    }
  )
)
```

الآن سنقوم بتوريث الصنف السابق ونعدل على الطريقة inc_age في الصنف الموروث ونضيف عليها اخباراً للتأكد من أن العمر الجديد ليس سالباً كما يلي:

التعليمة

```
InheritedStudent <- setRefClass("InheritedStudent",
  fields=list(country="character"),
  contains="student",
  methods=list(
    dec_age = function(x) {
      if((age - x)<0)  stop("Age cannot be negative")
      age <<- age - x
    }
  )
)

s <- InheritedStudent (name="John", age=21, avg=95, country="France")
s$dec_age(5)
s$age
s$dec_age(20)
s$age
```

الناتج

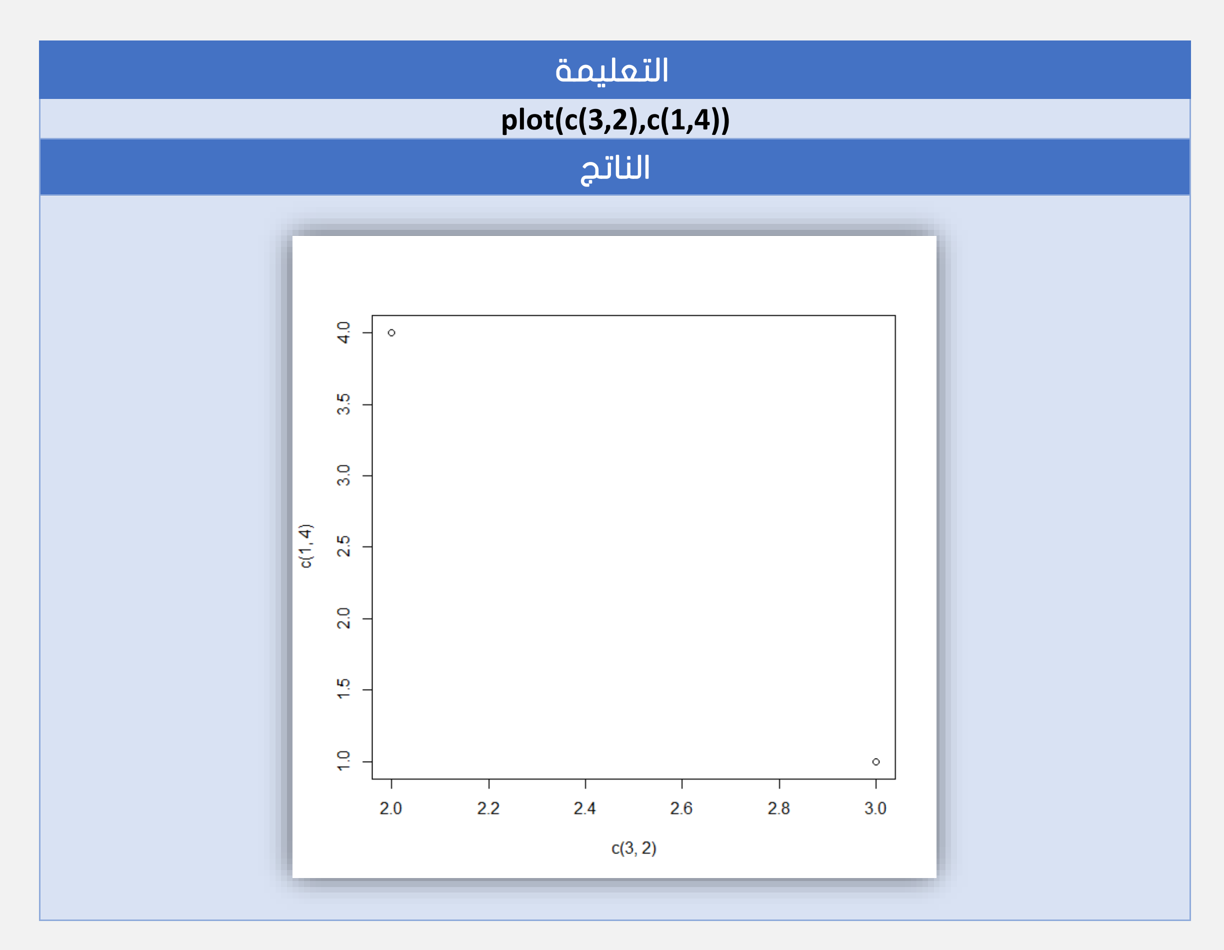
```
[1] 16
Error in s$dec_age(20) : Age cannot be negative
[1] 16
```

الفصل الخامس
الرسم والمخططات
(Graphs and Charts)

التابع plot:

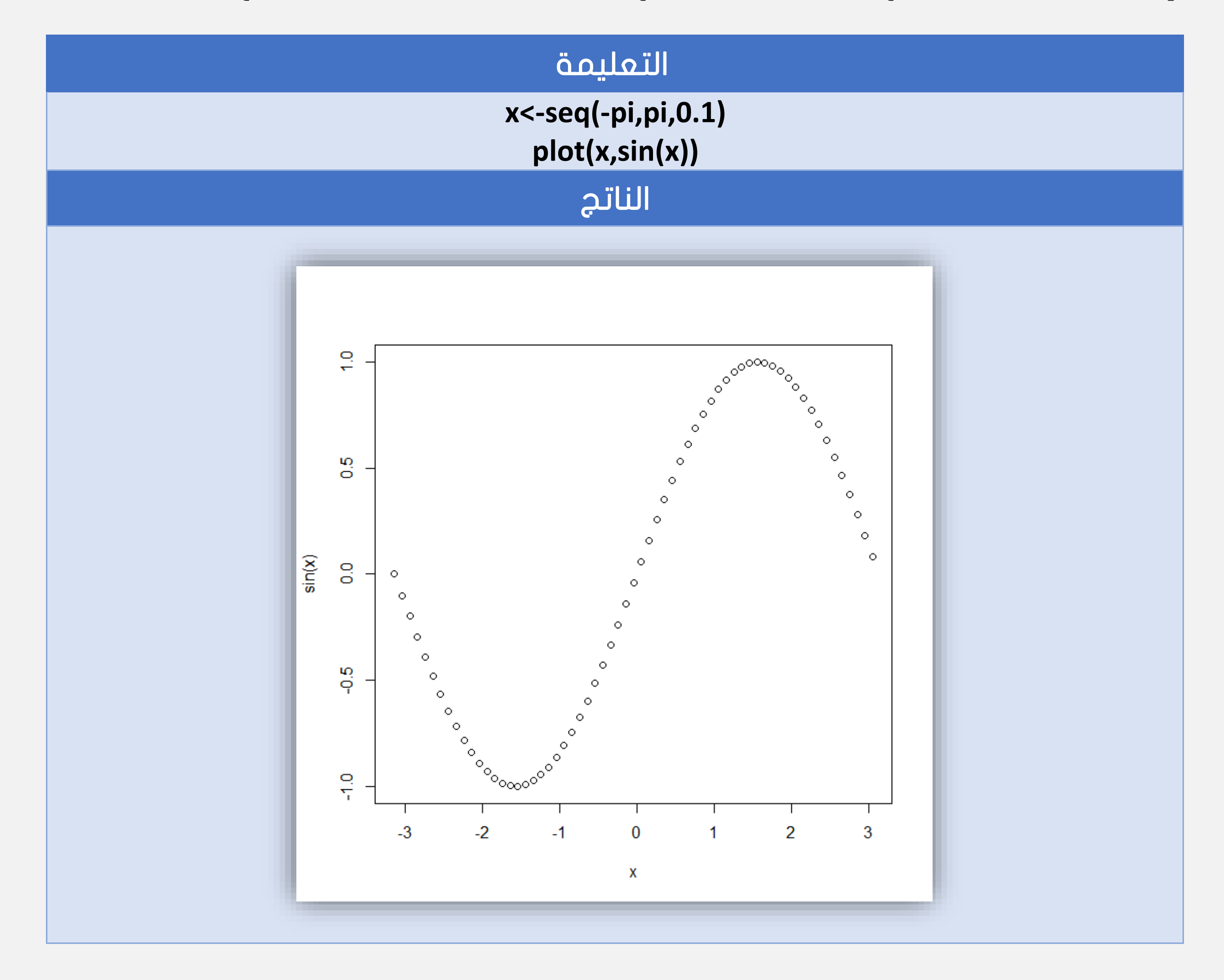
هو من أكثر توابع الرسم استخداماً في R وهو تابع عام (Generic function)، أي أنه يمتلك العديد من الطرائق حتى يلائم الكائنات الممررة له.

أبسط استخدام للتابع plot هو عندما نمرر له شعاعاً فيقوم برسم مخطط انتشار له، فمثلاً:



التعليمة
plot(c(3,2),c(1,4))
الناتج

بإمكاننا استخدام التابع ()plot لرسم توابع رياضية أيضاً، فمثلاً لرسم التابع sin(x) نكتب:



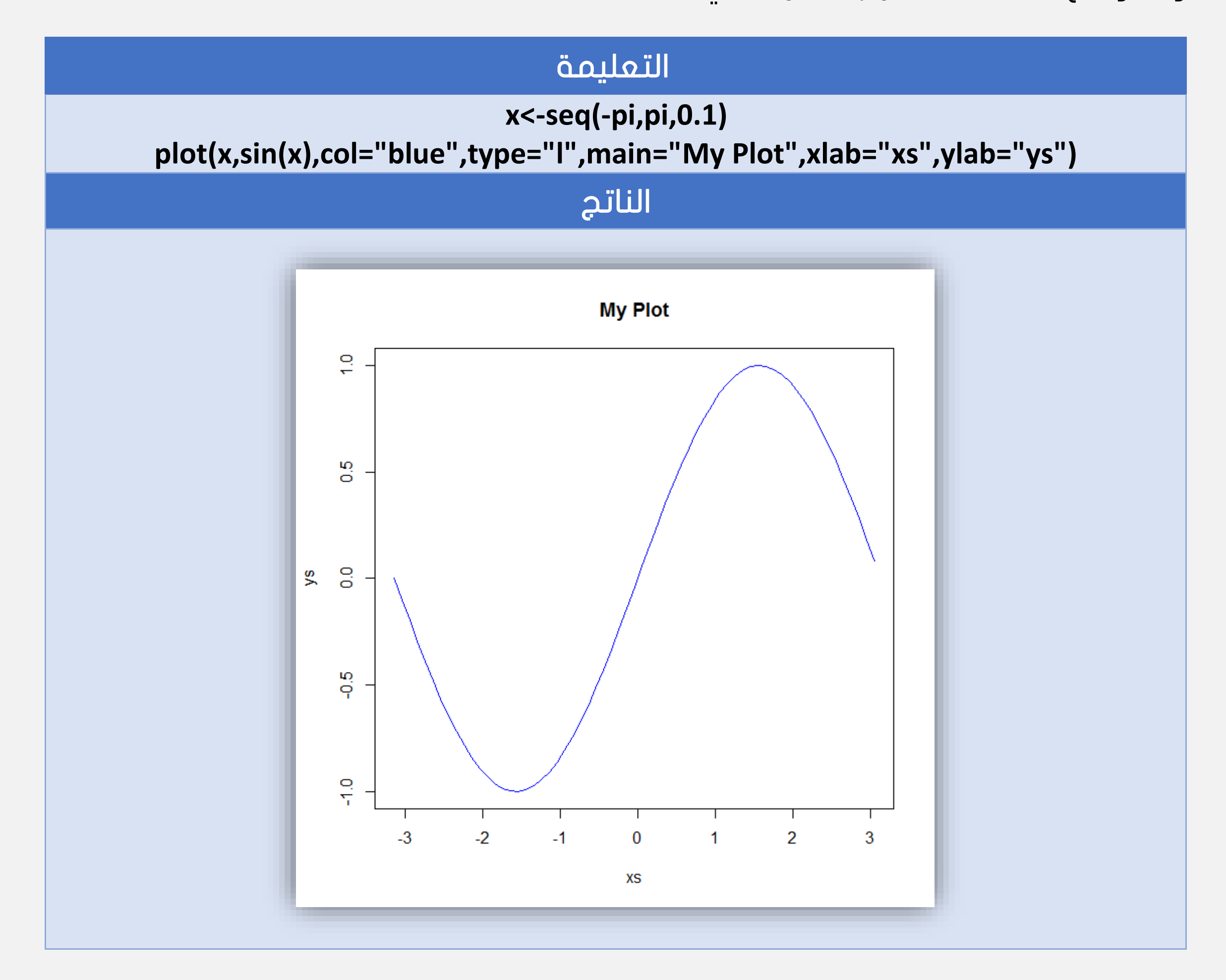
التعليمة
x<-seq(-pi,pi,0.1) plot(x,sin(x))
الناتج

بعض خصائص التابع plot:

الجدول الآتي يبين أهم الخصائص التي يمكن استخدامها مع التابع plot:

التعليمة	الوظيفة	ملاحظات
main="text"	لوضع عنوان عام للرسم مكان text	
xlab="text"	لوضع تسمية للمحور x`x مكان text	
ylab="text"	لوضع تسمية للمحور y`y مكان text	
type="tp"	لتغيير نمط الرسم إلى tp.	من الأنماط: p للنقاط، l للمستقيم، b نقاط ومستقيمات، ...
col="cl"	لتغيير لون الرسم إلى cl.	من الألوان: blue,red,yellow,...

وسنوضح هذه الخصائص بالمثال الآتي:

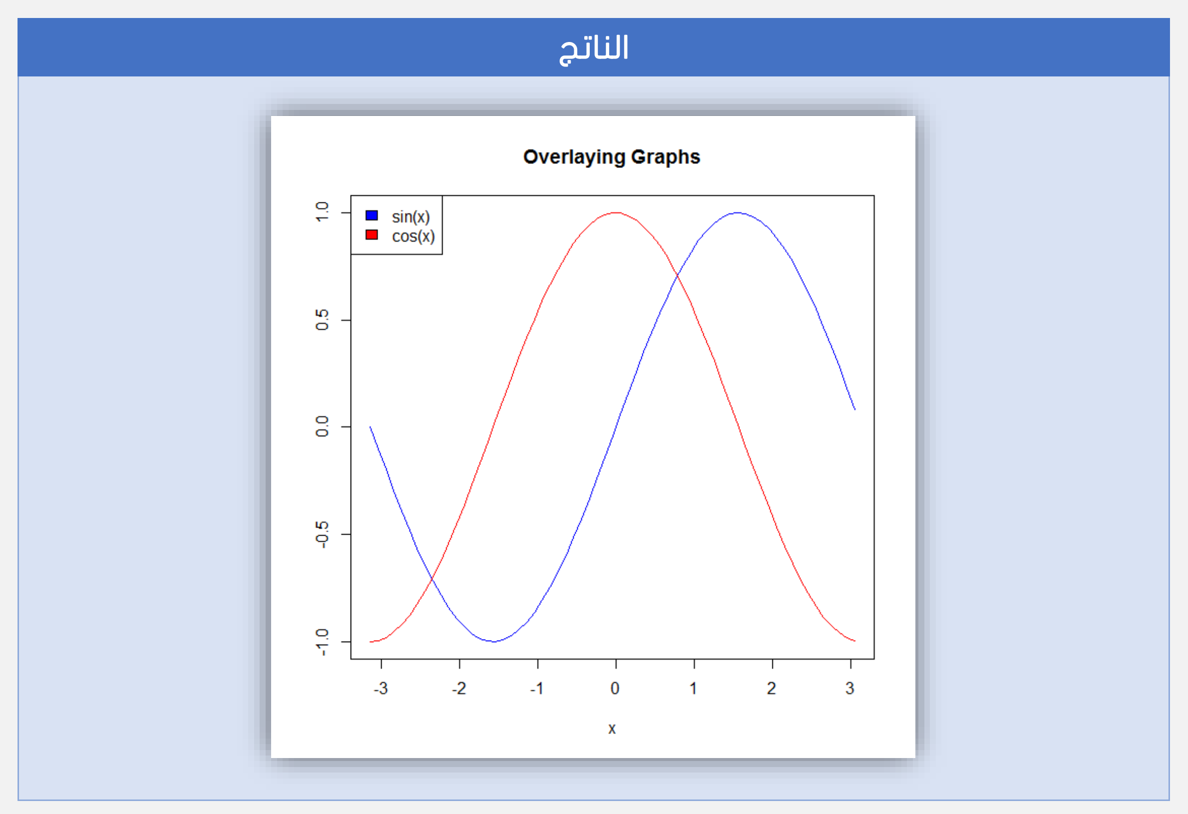
التعليمة
x<-seq(-pi,pi,0.1) plot(x,sin(x),col="blue",type="l",main="My Plot",xlab="xs",ylab="ys")
الناتج


رسم عدة توابع في نفس النافذة (Overlaying Plots):

عند كل استدعاء للتابع plot إن الرسم القديم ستتم إزالته ثم إدراج رسم جديد، أما إذا أردنا إدراج أكثر من رسمة في نفس النافذة فنستخدم التابع lines أو التابع points، والمثال الآتي يقوم برسم: كل من تابع sin وتابع cos في نفس النافذة باستخدام التعليمة plot أولاً لرسم التابع sin ثم التعليمة lines ثانياً لرسم التابع cos:

التعليمة
x<-seq(-pi,pi,0.1) plot(x, sin(x),main="Overlaying Graphs",ylab="",type="l",col="blue") lines(x,cos(x), col="red") legend("topleft",c("sin(x)","cos(x)"),fill=c("blue","red"))

الناتج



قمنا باستخدام التابع legend والذي يمثل مفتاح الرسم البياني والذي تم وضعه في أعلى أيسر الرسم كوننا وضعنا الخاصية "topleft".

رسم عدة مخططات متجاورة في نفس النافذة (Subplots):

نقصد تقسيم ساحة العمل إلى m سطراً و n عموداً ثم رسم مخطط محدد في كل ساحة عمل جزئية، وهذا يتم باستخدام التابع ()par وبالاستعانة بالخاصية mfrow ولهذه العملية الشكل الآتي:

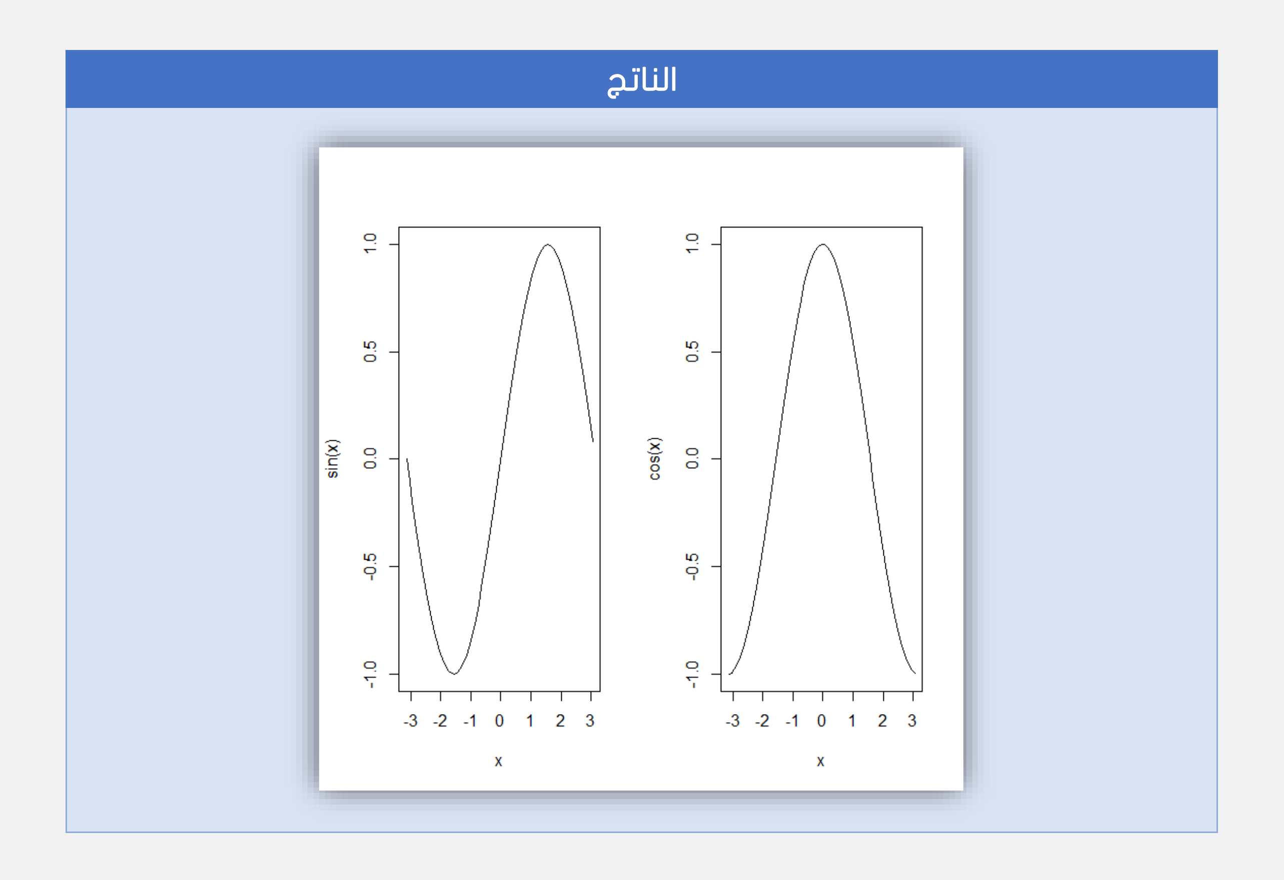
```
par(mfrow=c(m,n))
```

فلو أردنا رسم التابعين sin و cos بجانب بعضهما البعض نكتب:

التعليمة

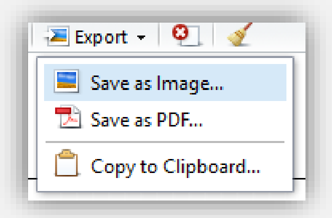
```
x<-seq(-pi,pi,0.1)
par(mfrow=c(1,2))
plot(x,sin(x),type="l")
plot(x,cos(x),type="l")
```

الناتج



نسخ الرسوم (Copying Plots):

عند النقر على الزر export تظهر لنا الخيارات الآتية:



الخيار copy to Clipboard يقوم بنسخ الرسم البياني وبإمكاننا بعد ذلك لصق الرسم أينما نريد.

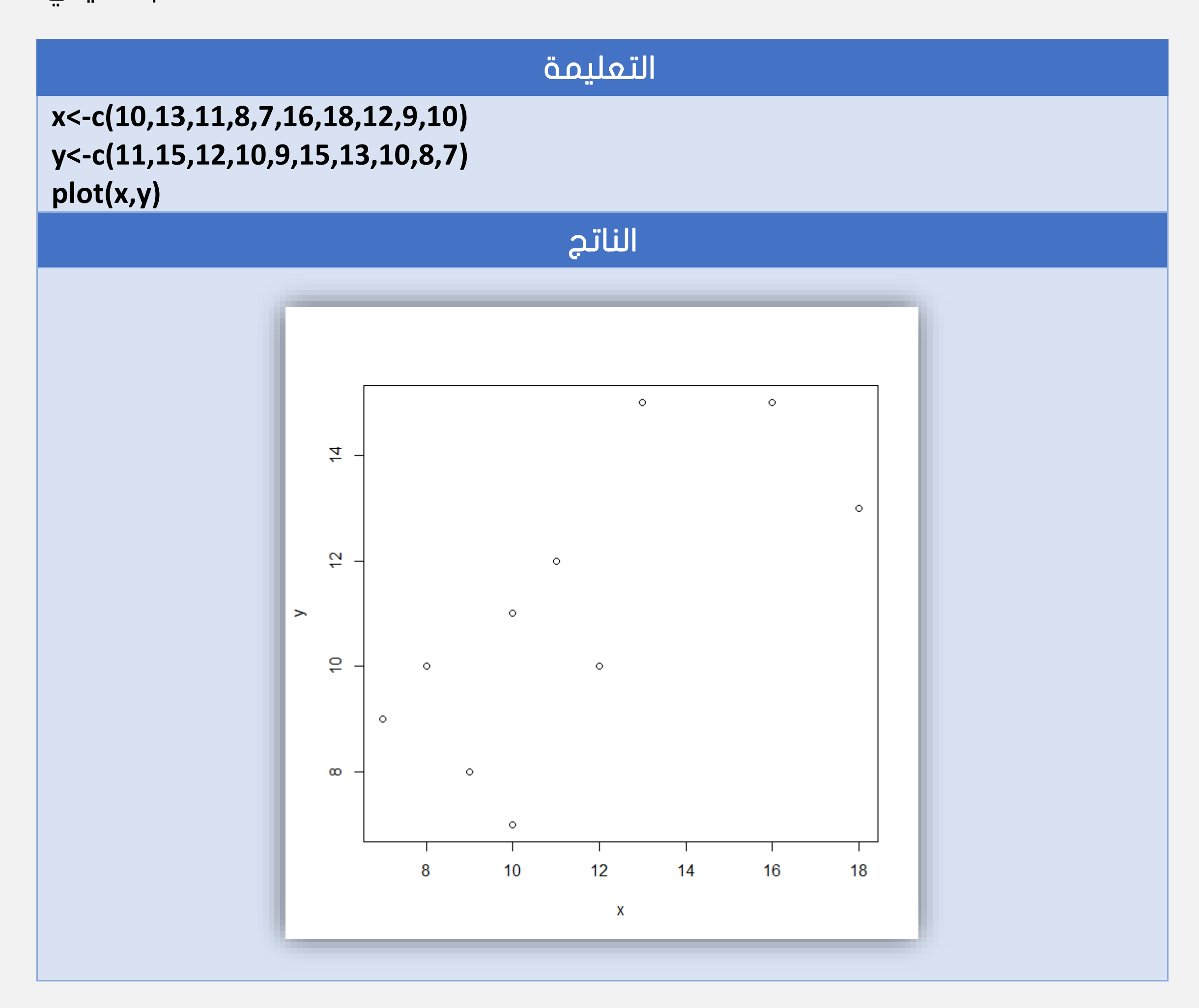
الخيار Save as Image يسمح لنا بحفظ الرسم البياني على القرص الصلب أينما نريد.

مخطط الانتشار (Scatter Plot):

يمكن رسم مخطط الانتشار لشعاعين x و y بالاستعانة بالتابع plot الذي تكلمنا عنه سابقاً، فلو أردنا رسم مخطط الانتشار للبيانات الآتية:

X	10	13	11	8	7	16	18	12	9	10
y	11	15	12	10	9	15	13	10	8	7

نكتب ما يلي:



التعليمة

```
x<-c(10,13,11,8,7,16,18,12,9,10)
y<-c(11,15,12,10,9,15,13,10,8,7)
plot(x,y)
```

الناتج

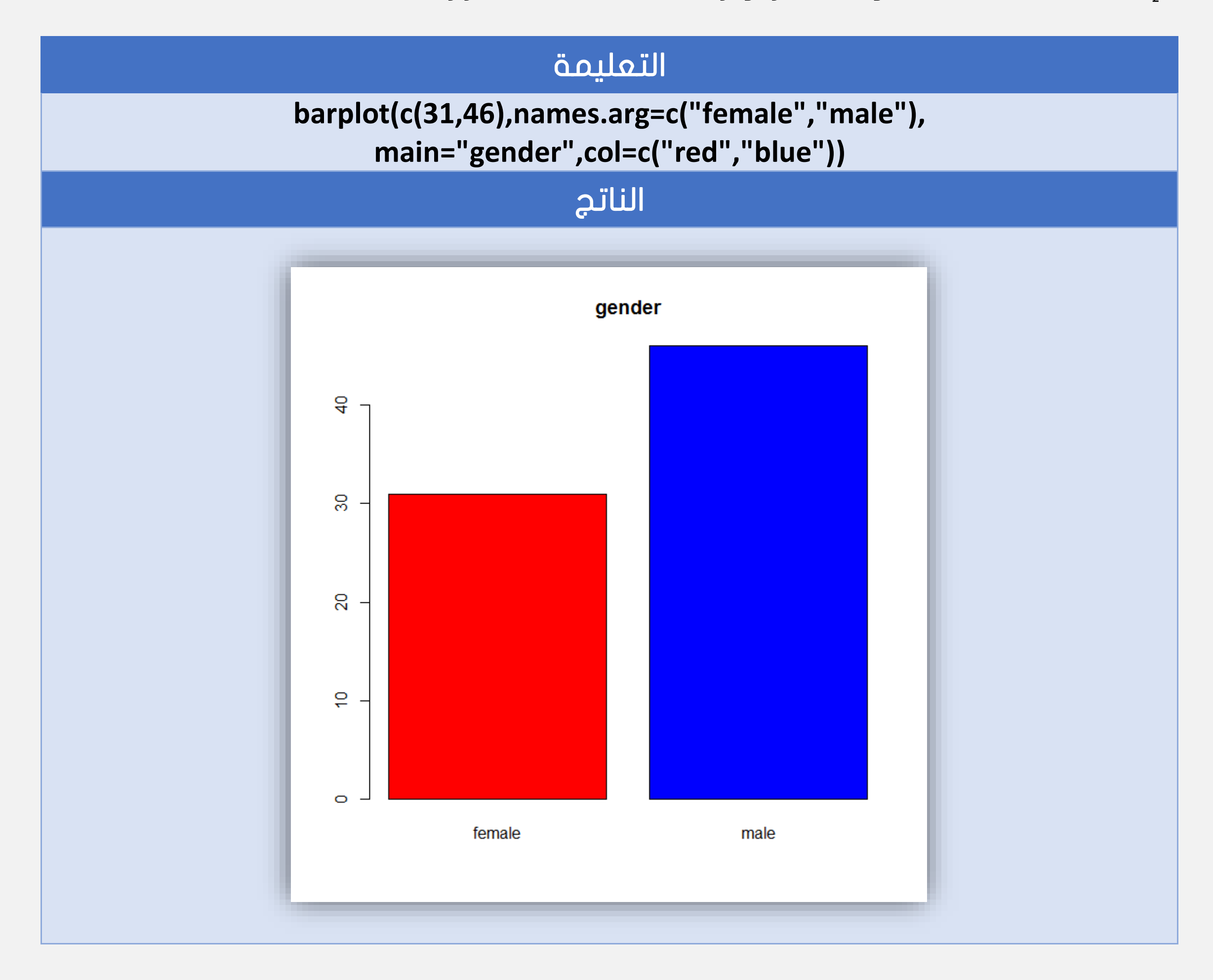
مخطط الأعمدة (Bar Plot):

يمكن رسم مخطط الأعمدة لشعاع التكرارات أو النسب f بالشكل:

```
barplot(f,xlab="xlab",ylab="ylab",main="Title",names.arg=names)
```

حيث تستخدم الخاصية names.arg لوضع تسمية توضيحية لما تمثله التكرارات f على المحور ox على شكل شعاع names.

فإذا كان لدينا 31 أنثى و 46 ذكراً وأردنا تمثيل هذه التكرارات بمخطط الأعمدة نكتب:

التعليمة
barplot(c(31,46),names.arg=c("female","male"), main="gender",col=c("red","blue"))
الناتج


ملاحظة:

يمكننا قلب الأعمدة ورسمها بشكل أفقي باستخدام الخاصية horiz=TRUE.

ملاحظة:

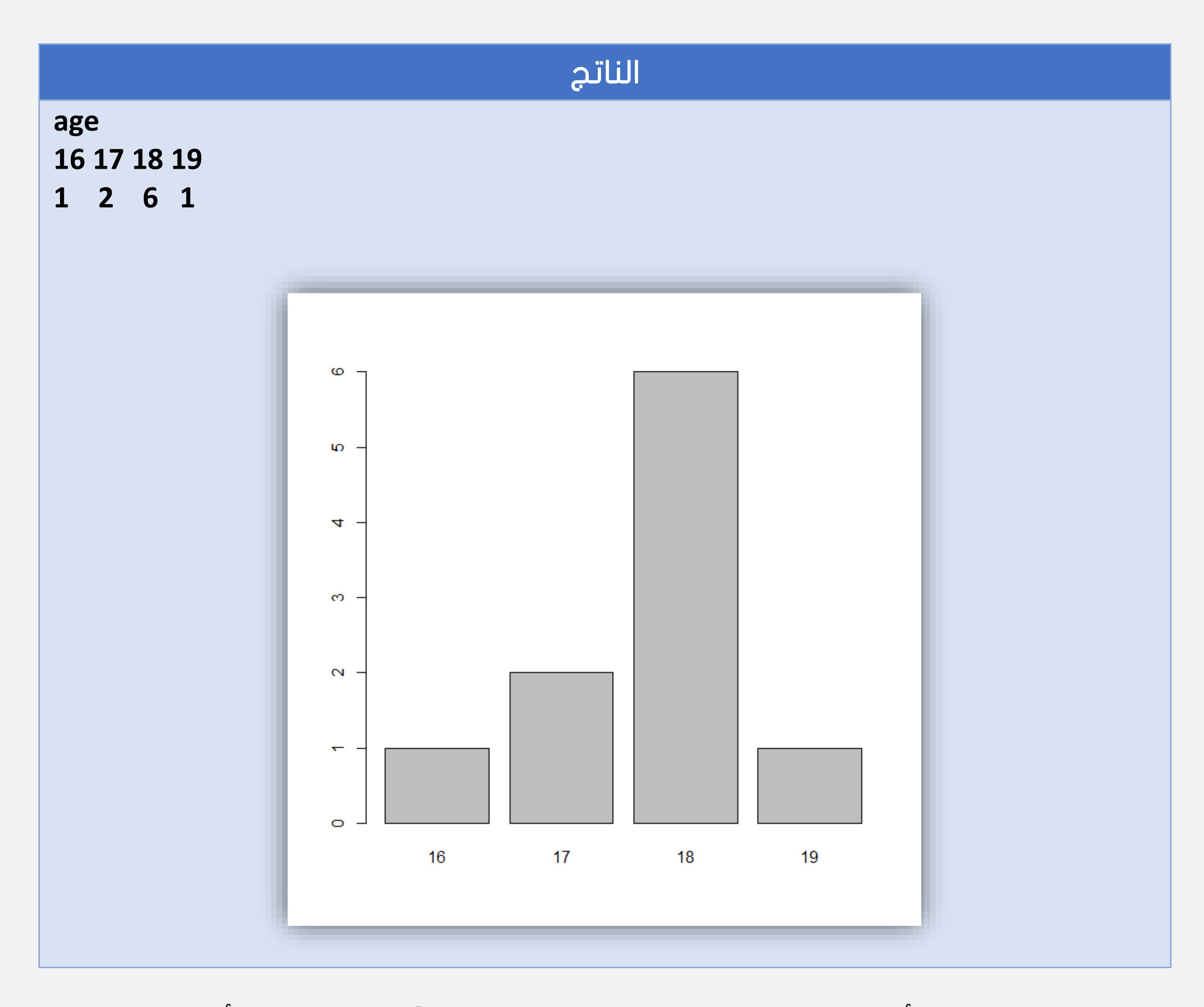
لو كانت لدينا عدة بيانات وأردنا رسم مخطط الأعمدة لتكراراتها نحولها أولاً لجدول تكراري باستخدام التعليمة table ثم نرسمها.

مثال

```
age <- c(17,18,18,17,18,19,18,16,18,18)
table(age)
barplot(table(age))
```

الناتج

```
age
16 17 18 19
1   2   6  1
```



يمكننا رسم مخطط أعمدة مزدوج (العمر حسب الجنس مثلاً) وذلك بشرط أن تكون بياناتنا على شكل جدول table، وسنوضح ذلك في المثال الآتي:

مثال

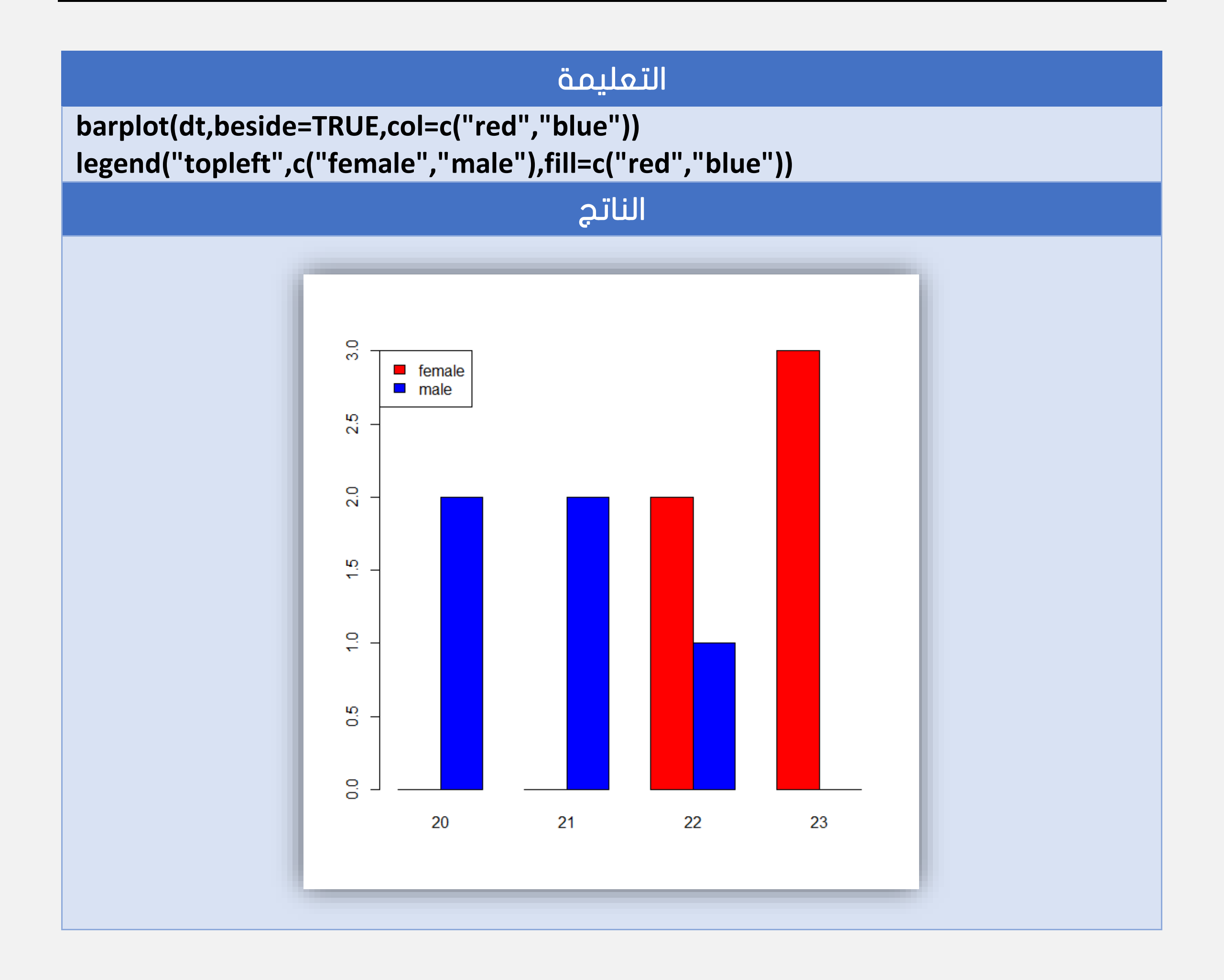
```
data<-
data.frame("gender"=c(rep("m",5),rep("f",5)),"age"=c(rep(20:21,2),rep(22:23,3)))
dt<-table(data)
dt
```

الناتج

```
      age
gender 20 21 22 23
    f      0  0  2  3
    m     2  2  1  0
```

قمنا بالتعليمة السابقة بتشكيل جدول يمثل أعمار افتراضية لذكور وإناث، وسنقوم الآن برسم مخطط الأعمدة الموافق له بالشكل الآتي:

التعليمة
barplot(dt,beside=TRUE,col=c("red","blue")) legend("topleft",c("female","male"),fill=c("red","blue"))
الناتج



مخطط الفطيرة (Pie Chart):

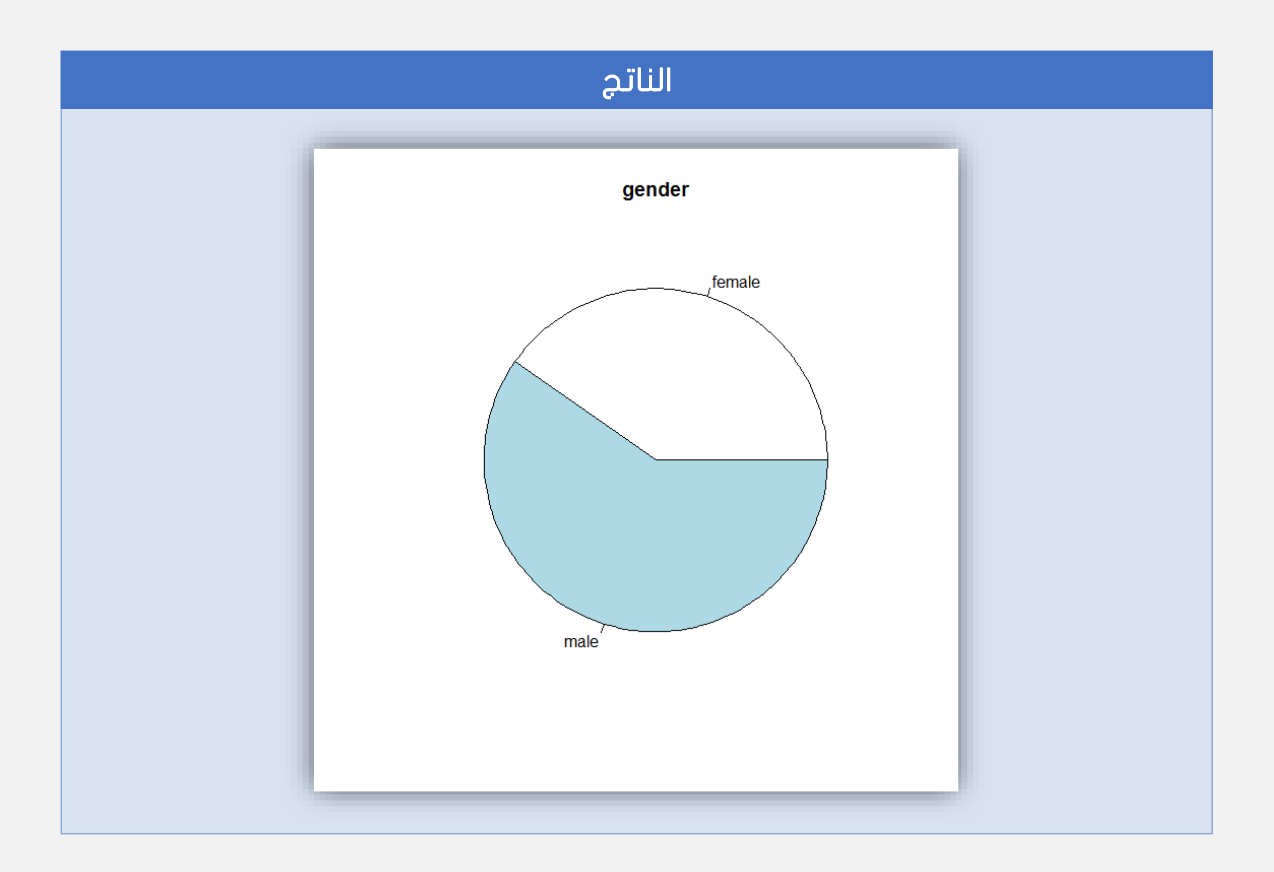
يمكن رسم مخطط الفطيرة لشعاع التكرارات أو النسب f بالشكل:

```
pie(f ,main=”Title”,labels=names)
```

تمثل labels أسماء ووصف التكرارات f، ولرسم نفس المثال السابق بمخطط الفطيرة نكتب:

التعليمة
pie(c(31,46),labels=c("female","male"),main="gender")

الناتج



مخطط الصندوق (Box Plot):

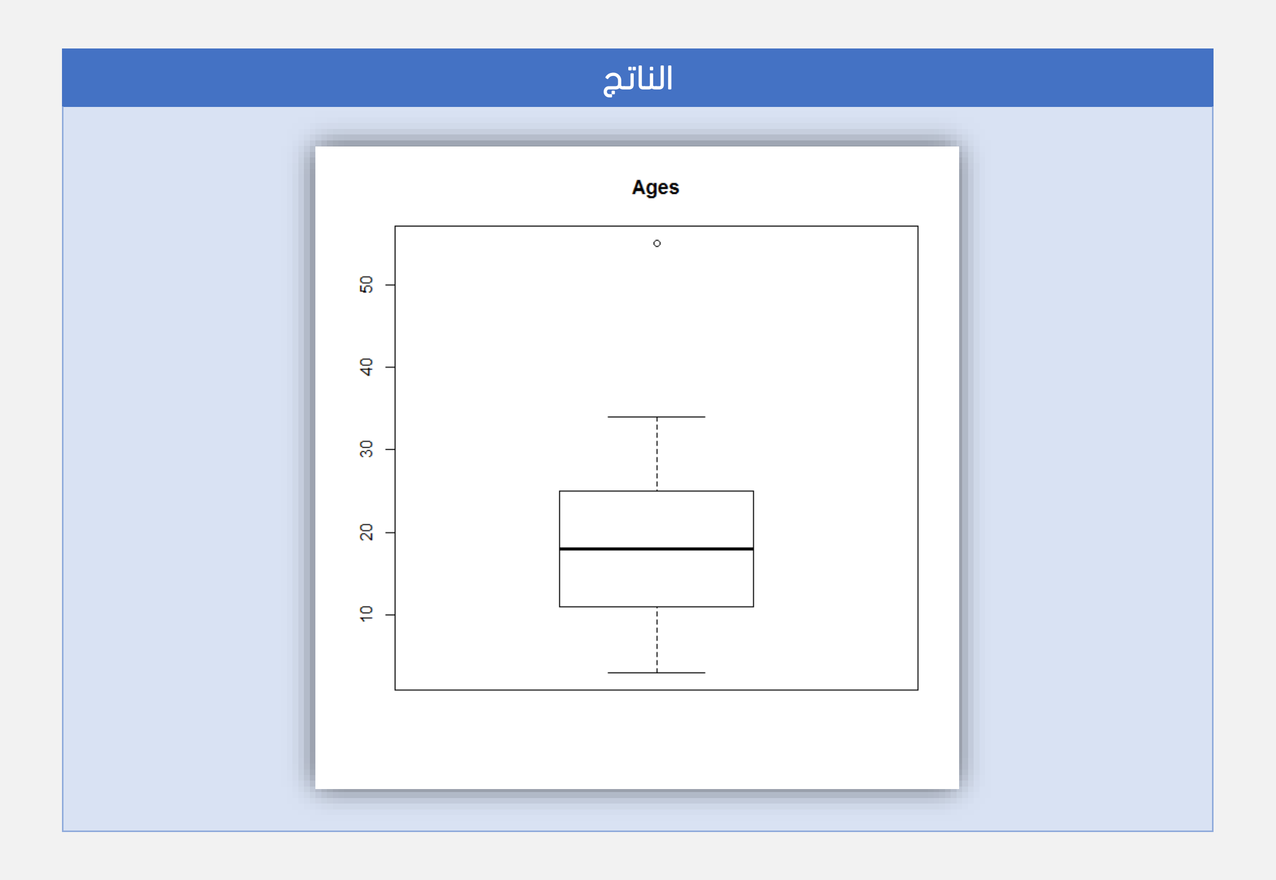
يمكن استخدام مخطط الصندوق لتمثيل شعاع البيانات الكمية x وذلك بالشكل:

```
boxplot(x,ylab="ylab",main="Title")
```

مثال

```
age <- c(10,7,18,10,12,19,18,20,3,14,19,55,25,31,26,11,34)
boxplot(age,main="Ages")
```

الناتج



للمخطط السابق فائدة وصفية كبيرة، حيث تمثل حافة الصندوق الدنيا الربيع الأول، وتمثل حافة الصندوق العليا الربيع الثالث، بينما يمثل الخط الذي هو داخل الصندوق الربيع الثاني، وبالتالي يحوي الصندوق 50% من البيانات، كما أن الدوائر التي قد نراها خارج الصندوق تمثل القيم الشاذة.

ملاحظة:

يمكن رسم أكثر من مخطط في نفس النافذة بوضع الأشعة المراد رسمها بجانب بعضها البعض مفصولاً بينها بفواصل، مثلاً:

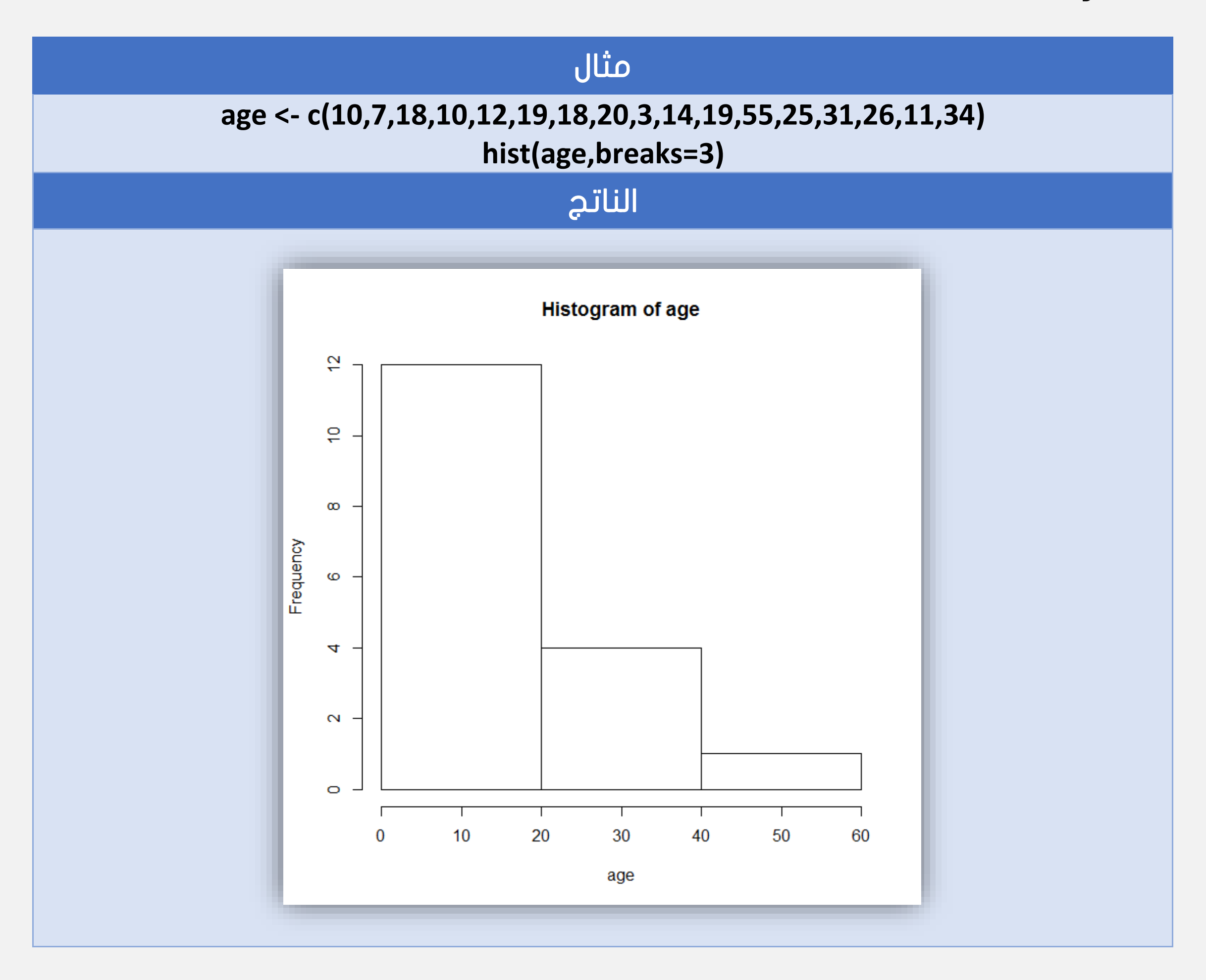
```
boxplot(age1,age2)
```

المدرج التكراري (Histogram):

يستخدم مع البيانات الكمية لرسم الشعاع x بالشكل:

```
hist(x,breaks=k-1)
```

تمثل breaks عدد النقاط التي نريد استخدامها لتجزئة البيانات، وهو أيضاً يمثل عدد الفئات التي نريد أن نقسم البيانات لها ناقصاً واحد، وبالتالي هو يمثل عدد الأعمدة التي ستظهر ناقصاً واحد.



الفصل السادس
نظرية الاحتمالات
(Probability Theory)

قبل أن تبدأ بهذا الفصل عليك أولاً أن تقوم بتحميل حزمة الاحتمالات prob من موقع CRAN أو باستخدام التعليمة الآتية:

```
install.packages("prob")
```

ثم تقوم بتحميل مكتبة الاحتمالات بالشكل:

```
library("prob")
```

نمذجة فضاء العينة لبعض التجارب الاحتمالية (Sample Spaces):

أحد أهم التجارب الاحتمالية التي قد نهتم بها هي **تجربة رمي قطعة نقود متزنة n مرة**، وحتى نستخلص النتائج المرجوة من هذه التجربة علينا أولاً معرفة فضاء الحالة، وهذا ما يقدمه لنا R وبسهولة مطلقة باستخدام التعليمة:

```
tosscoin(n)
```

فمثلاً عند رمي قطعة نقود متزنة 3 مرات إن فضاء الحالة سيكون:

مثال
tosscoin(3)
الناتج

	toss1	toss2	toss3
1	H	H	H
2	T	H	H
3	H	T	H
4	T	T	H
5	H	H	T
6	T	H	T
7	H	T	T
8	T	T	T

الجدول السابق يبين أنه عند رمي قطعة النقود 3 مرات فإن عدد الحالات التي قد نصادفها هو 8 حالات، كل حالة من هذه الحالات ممثلة بسطر مستقل.

تجربة أخرى تصادفنا كثيراً في كتب نظرية الاحتمالات ألا وهي **تجربة رمي حجر نرد متزن n مرة**، وأيضاً يتيح لنا R إيجاد فضاء الحالة لهذه التجربة باستخدام التعليمة:

```
rolldie(n)
```

ومن التجارب التي نصادفها أيضاً هي **تجربة سحب k كرة من صندوق أو وعاء كبير (Urn) فيه كرات مرقمة من 1 إلى n** والتي يمكن أيضاً نمذجتها باستخدام R بالشكل الآتي:

```
urnsamples(1:n,size=k,replace=TRUE or FALSE,ordered = TRUE or FALSE)
```

حيث يمثل الوسيط الأول 1:n ترقيم الكرات والتي عددها n، أما الوسيط الثاني size=k فيمثل عدد الكرات المراد سحبها، ويمثل الوسيط الثالث replace = TRUE or false حالة السحب هل هي مع إعادة (TRUE) أو بدون إعادة، أما الوسيط الأخير ordered=TRUE or FALSE فيمثل كون الترتيب مهماً (TRUE) أو غير مهم (FALSE).

فلو أردنا سحب كرتين من ثلاث كرات على الترتيب وبدون إعادة يكون:

مثال
urnsamples(1:3,size = 2,replace = FALSE, ordered = TRUE)
الناتج

```
  X1 X2
1  1  2
2  2  1
3  1  3
4  3  1
5  2  3
6  3  2
```

وآخر تجربة شهيرة نصادفها في مسائل الاحتمالات هي تجربة أوراق اللعب، ويمثل فضاء العينة لأوراق اللعب بالتعليمة الآتية:

```
cards()
```

الأحداث (Events):

نستطيع تعريف الحدث على أنه مجموعة جزئية من فضاء العينة، فلو كانت لدينا التجربة:

مثال
S<-tosscoin(2,makespace=TRUE) S
الناتج
toss1 toss2 probs 1 H H 0.25 2 T H 0.25 3 H T 0.25 4 T T 0.25

فيمكننا أخذ الحدث A الممثل بأول ثلاثة سطور بالشكل:

التعليمة
S[1:3,]
الناتج
toss1 toss2 probs 1 H H 0.25 2 T H 0.25 3 H T 0.25

أو أخذ السطر الثاني والرابع مثلاً بالشكل:

التعليمة
S[c(2,4,]
الناتج
toss1 toss2 probs 2 T H 0.25 4 T T 0.25

ملاحظة:

الخاصية makespace=TRUE تجعل R يقوم بحساب الاحتمالات لفضاء العينة المولد.

ملاحظة:

بإمكاننا تعريف الفضاء الاحتمالي واحتمالاته كما نشاء، والمثال الآتي يعرف احتمالات رمي قطعة نقود غير متجانسة مرة واحدة:

probspace(tosscoin(1),probs=c(0.3,0.7))

بإمكاننا أيضاً اختيار الأسطر التي تمثل شرطاً محدداً باستخدام التابع subset فمثلاً إن الحدث الذي يمثل أوراق لعب الكوبة يمكن تمثيله بالشكل:

مثال

```
S<-cards()
subset(S,suit=="Heart")
```

الناتج

```
     rank   suit
27    2     Heart
28    3     Heart
29    4     Heart
30    5     Heart
31    6     Heart
32    7     Heart
33    8     Heart
34    9     Heart
35   10     Heart
36    J     Heart
37    Q     Heart
38    K     Heart
39    A     Heart
```

أما أوراق اللعب التي تحمل أرقاماً بين 7 و 9 فيمكن تمثيلها بالشكل:

مثال

```
S<-cards()
subset(S,rank %in% 7:9)
```

الناتج

```
     rank  suit
6     7     Club
7     8     Club
8     9     Club
19    7     Diamond
20    8     Diamond
21    9     Diamond
32    7     Heart
33    8     Heart
34    9     Heart
45    7     Spade
46    8     Spade
47    9     Spade
```

في تجربة إلقاء حجر نرد ثلاث مرات، إن حدث كون مجموع الوجوه الثلاثة أكبر من 16 يمثل بالشكل الآتي:

مثال
subset(rolldie(3), X1 + X2 + X3 > 16)

الناتج			
	X1	X2	X3
180	6	6	5
210	6	5	6
215	5	6	6
216	6	6	6

التابع %in% والتابع isin:

يستخدم التابع %in% لمعرفة فيما إذا كانت كل قيمة من الشعاع y تقع في شعاع آخر x:

مثال

```
x<-1:10
y<-8:12
y %in% x
```

الناتج

```
TRUE  TRUE  TRUE FALSE FALSE
```

ذلك لأن كل من 8 و 9 و 10 تقع في الشعاع x لكن 11 و 12 لا تقعان فيه.

يستخدم التابع isin لمعرفة إذا كان الشعاع y بأكمله يقع في شعاع آخر x:

مثال

```
x<-1:10
y<-8:12
isin(y,x)
```

الناتج

```
FALSE
```

الاجتماع والتقاطع والفرق (Union, Intersection and Difference):

الاسم	الرمز	التعريف حسب العناصر	التعليمة في R
الاجتماع	$A \cup B$	في A أو B أو كلاهما	union(A,B)
التقاطع	$A \cap B$	في A و B معاً	intersect(A,B)
الفرق	$A \backslash B$	في A وليس في B	setdiff(A,B)

بإمكاننا الاستفادة من التعليمات السابقة للتعامل مع الأحداث، فمثلاً لو أردنا تعريف الحدث الذي يقع إذا كان مجموع الرقمين الظاهرين على حجري نرد أكبر من 8 وكان الوجهان فرديين نكتب ما يلي:

مثال

```
A<-subset(rolldie(2),X1+X2>8)
B<-subset(rolldie(2),X1 %% 2 == 0 && X2 && 2 == 0)
intersect(A,B)
```

الناتج

```
     X1  X2
24    6   4
34    4   6
36    6   6
```

<u>مسألة:</u>

في تجربة إلقاء حجري نرد معاً احسب احتمال الحصول على وجهين مجموعهما يقبل القسمة على 3.

الحل

```
A<-subset(rolldie(2,makespace=TRUE),(X1+X2) %% 3 == 0)
Prob(A)
```

الناتج

```
0.3333333
```

الاحتمالات الشرطية (Conditional Probabilities):

يمكننا حساب الاحتمال الشرطي P(A|B) في R باستخدام التعليمة:

```
Prob(A,given=B)
```

فلو أردنا حساب احتمال كون الورقة المسحوبة تحمل العدد 5 علماً أنها كوبة نكتب:

الحل

```
A<-subset(cards(makespace=TRUE),suit=="Heart")
B<-subset(cards(makespace=TRUE),rank == 5)
Prob(B,given=A)
```

الناتج

```
0.07692308
```

<u>مسألة:</u> صندوق يحتوي سبع كرات حمراوات وثلاث كرات خضراوات، قمنا بسحب ثلاث كرات على التتالي بدون إعادة، والمطلوب: احسب احتمال أن تكون جميع الكرات المسحوبة حمراوات.

الحل

```
L <- rep(c("red", "green"), times = c(7, 3))
M <- urnsamples(L, size = 3, replace = FALSE, ordered = TRUE)
N <- probspace(M)
Prob(N, isrep(N, "red", 3))
```

الناتج

```
0.2916667
```

ما احتمال الحصول على كرة حمراء ثم كرة خضراء ثم كرة حمراء؟

الحل

```
Prob(N, isin(N, c("red", "green", "red"), ordered = TRUE))
```

الناتج

```
0.175
```

ما احتمال الحصول على كرتين حمراوين وكرة خضراء؟

الحل

```
Prob(N, isin(N, c("red", "green", "red")))
```

الناتج

```
0.525
```

التوزيع الثنائي $B(n,p)$ (Binomial Distribution):

تعطى دالة الاحتمال للتوزيع الثنائي بالعلاقة:

$$P(X=k)=C_k^n\,p^k\,q^{n-k}\;;k=0,1,\dots,n$$

ونعبر عن ذلك الاحتمال في R بالشكل:

```
dbinom(k,n,p)
```

أما دالة التوزيع المتجمعة $F(k)=P(X\leq k)$ فنعرفها في R بالشكل:

```
pbinom(k,n,p)
```

بإمكاننا إيجاد العدد a والذي يحقق $F(a)=prob$ باستخدام التعليمة:

```
qbinom(prob,n,p)
```

كما بإمكاننا توليد عينة عشوائية حجمها N تخضع للتوزيع الثنائي $B(n,p)$ باستخدام التعليمة:

```
rbinom(N,n,p)
```

توزيع بواسون $poi(\lambda)$ (Poisson Distribution):

تعطى دالة الاحتمال لتوزيع بواسون بالعلاقة:

$$P(X=k)=\frac{\lambda^k}{k!}e^{-\lambda}\;;k=0,1,\dots$$

ونعبر عن ذلك الاحتمال في R بالشكل:

```
dpois(k,lambda)
```

أما دالة التوزيع المتجمعة $F(k)=P(X\leq k)$ فنعرفها في R بالشكل:

```
ppois(k,lambda)
```

بإمكاننا إيجاد العدد a والذي يحقق $F(a)=prob$ باستخدام التعليمة:

```
qpois(prob,lambda)
```

كما بإمكاننا توليد عينة عشوائية حجمها N تخضع للتوزيع بواسون $Poi(\lambda)$ باستخدام التعليمة:

```
rpois(N,lambda)
```

التوزيع المنتظم $U(a,b)$ (Uniform Distribution):

تعطى دالة الكثافة الاحتمالية للتوزيع المنتظم بالعلاقة:

$$f(x) = \frac{1}{b-a}; \mathrm{a} < x < \mathrm{b}$$

ونعبر عنها في R بالشكل:

```
dunif(x,a,b)
```

أما دالة التوزيع المتجمعة $F(k) = P(X \leq k)$ فنعرفها في R بالشكل:

```
punif(x,a,b)
```

بإمكاننا إيجاد العدد a والذي يحقق $F(a) = prob$ باستخدام التعليمة:

```
qunif(x,a,b)
```

كما بإمكاننا توليد عينة عشوائية حجمها N تخضع للتوزيع المنتظم $U(a,b)$ باستخدام التعليمة:

```
runif(N,a,b)
```

التوزيع الأسي $exp(\lambda)$ (Exponential Distribution):

تعطى دالة الكثافة الاحتمالية للتوزيع الأسي بالعلاقة:

$$f(x) = \lambda e^{-\lambda x}; \mathrm{x} > 0$$

ونعبر عنها في R بالشكل:

```
dexp(x,lambda)
```

أما دالة التوزيع المتجمعة $F(k) = P(X \leq k)$ فنعرفها في R بالشكل:

```
pexp(x,lambda)
```

بإمكاننا إيجاد العدد a والذي يحقق $F(a) = prob$ باستخدام التعليمة:

```
qexp(prob,lambda)
```

كما بإمكاننا توليد عينة عشوائية حجمها N تخضع للتوزيع الأسي $exp(\lambda)$ باستخدام التعليمة:

```
rexp(N,lambda)
```

التوزيع الطبيعي $N(\mu,\sigma^2)$ (Normal Distribution):

تعطى دالة الكثافة الاحتمالية للتوزيع الطبيعي بالعلاقة:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

ونعبر عنها في R بالشكل:

```
dnorm(x,mu,sigma)
```

أما دالة التوزيع المتجمعة $F(k) = P(X \leq k)$ فنعرفها في R بالشكل:

```
pnorm(x,mu,sigma)
```

بإمكاننا إيجاد العدد a والذي يحقق $F(a) = prob$ باستخدام التعليمة:

```
qnorm(prob,mu,sigma)
```

بإمكاننا توليد عينة عشوائية حجمها N تخضع للتوزيع الطبيعي $N(\mu,\sigma^2)$ باستخدام التعليمة:

```
rnorm(N,mu,sigma)
```

توزيع كاي-مربع $\chi^2(n)$ (Chi-Square Distribution):

تعطى دالة الكثافة الاحتمالية لتوزيع كاي-مربع بالعلاقة:

$$f(x) = \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} e^{-\frac{x}{2}} x^{\frac{n}{2}-1}; x > 0$$

ونعبر عنها في R بالشكل:

```
dchisq(x,n)
```

أما دالة التوزيع المتجمعة $F(k) = P(X \leq k)$ فنعرفها في R بالشكل:

```
pchisq(x,n)
```

بإمكاننا إيجاد العدد a والذي يحقق $F(a) = prob$ باستخدام التعليمة:

```
qchisq(prob,n)
```

بإمكاننا توليد عينة عشوائية حجمها N تخضع لتوزيع $\chi^2(n)$ باستخدام التعليمة:

```
rchisq(N,n)
```

توزيع ستيودينت $t(n)$ (Student Distribution):

تعطى دالة الكثافة الاحتمالية لتوزيع ستيودينت بالعلاقة:

$$f(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}}$$

ونعبر عنها في R بالشكل:

```
dt(x,n)
```

أما دالة التوزيع المتجمعة $F(k) = P(X \leq k)$ فنعرفها في R بالشكل:

```
pt(x,n)
```

بإمكاننا إيجاد العدد a والذي يحقق $F(a) = prob$ باستخدام التعليمة:

```
qt(prob,n)
```

بإمكاننا توليد عينة عشوائية حجمها N تخضع للتوزيع $t(n)$ باستخدام التعليمة:

```
rt(N,n)
```

توزيع فيشر $F(m,n)$ (fisher Distribution):

تعطى دالة الكثافة الاحتمالية لتوزيع فيشر بالعلاقة:

$$f(x) = \frac{1}{\beta\left(\frac{m}{2},\frac{n}{2}\right)}\left(\frac{m}{n}\right)^{\frac{m}{2}} x^{\frac{m}{2}-1}\left(1+\frac{m}{n}x\right)^{-\frac{m+n}{2}} ; x > 0$$

ونعبر عنها في R بالشكل:

```
df(x,m,n)
```

أما دالة التوزيع المتجمعة $F(k) = P(X \leq k)$ فنعرفها في R بالشكل:

```
pf(x,m,n)
```

بإمكاننا إيجاد العدد a والذي يحقق $F(a) = prob$ باستخدام التعليمة:

```
qf(prob,m,n)
```

بإمكاننا توليد عينة عشوائية حجمها N تخضع للتوزيع $F(m,n)$ باستخدام التعليمة:

```
rf(N,m,n)
```

حساب التوقع الرياضي والتباين والانحراف المعياري لمتحول عشوائي (Expectation, Variance and Standard Deviation):

بإمكاننا حساب التوقع الرياضي لأي تابع g(X) للمتحول العشوائي X باستخدام التعليمة:

```
E(expression in X)
```

لكن أولاً علينا تحميل الحزمة distrEx بالشكل:

```
install.packages("distrEx")
```

فمثلاً إذا أردنا حساب E(2X+1) للتوزيع الطبيعي N(2,1) نكتب:

التعليمة
X<-Norm(2,1) E(2*X+1)
الناتج
5

كما يمكننا حساب التباين والانحراف المعياري باستخدام كل من التعليمتين:

```
var(expression in X)
sd(expression in X)
```

الفصل السابع
اختبارات الطبيعية واختبارات تجانس التباينات
(Normality Tests and Homogeneity of Variances Tests)

اختبار الفرضيات (Hypothesis Testing):

نستطيع تعريف الفرضية على أنها تخمين أو ادعاء يتعلق بوسطاء المجتمع الإحصائي وهي تحتمل الصحة والخطأ.

عندما رغب الباحث باختبار أي فرضية عليه أن يصوغها على شكل فرضيتين، فرضية تدعى فرضية العدم (Null Hypothesis) ويرمز لها بالرمز H_0 وفرضية تدعى الفرضية البديلة (Alternative Hypothesis) ويرمز لها بالرمز H_1، وهدف اختبار الفرضية هو دراسة إمكانية رفض الفرضية الابتدائية عند مستوى أهمية محدد، فبعد صياغة الفرضيتين الابتدائية والبديلة وتحديد مستوى الأهمية وعادة يكون 0.05 أو 0.01 أو 0.1، يقوم الباحث بإيجاد إحصاء الاختبار المناسب وحساب معنوية الاختبار (Significance) (P-Value)، ثم تأتي مرحلة اتخاذ القرار الإحصائي بمقارنة معنوية الاختبار مع مستوى الأهمية، فإذا كانت قيمة P-Value أقل من مستوى الأهمية نرفض الفرضية الابتدائية، وإذا كانت قيمة P-Value أكبر من مستوى الأهمية لا نستطيع رفض الفرضية الابتدائية.

بعض اختبارات الطبيعية (Some Normality Tests):

توجد العديد من الاختبارات التي يمكن بواسطتها التأكد من التزام البيانات بالتوزيع الطبيعي وسنحاول عرض أهمها:

اختبار Kolmogorov-Smirnov:

يستخدم لاختبار الفرضية:

$$H_0: X \sim N(\mu, \sigma^2)$$

مقابل:

$$H_1: X \nsim N(\mu, \sigma^2)$$

ويمكن تطبيقه باستخدام R بالشكل:

```
ks.test(X,”pnorm”,mu,sigma)
```

فإذا كانت P>0.05 فإن البيانات تتوزع وفق التوزيع الطبيعي.

مثال
x<-rnorm(300) ks.test(x,"pnorm")
الناتج
One-sample Kolmogorov-Smirnov test data: x D = 0.044407, p-value = 0.595 alternative hypothesis: two-sided

وبما أن P>0.05 فإن بياناتنا تخضع للتوزيع الطبيعي.

اختبار Shapiro-Wilk:

ويستخدم لاختبار الفرضية:

$$H_0: X \sim N(\mu, \sigma^2)$$

مقابل:

$$H_1: X \nsim N(\mu, \sigma^2)$$

ويمكن تطبيقه باستخدام R بالشكل:

```
shapiro.test(X)
```

فإذا كانت P>0.05 فإن البيانات تتوزع وفق التوزيع الطبيعي.

مخطط Q-Q (Q-Q Plot):

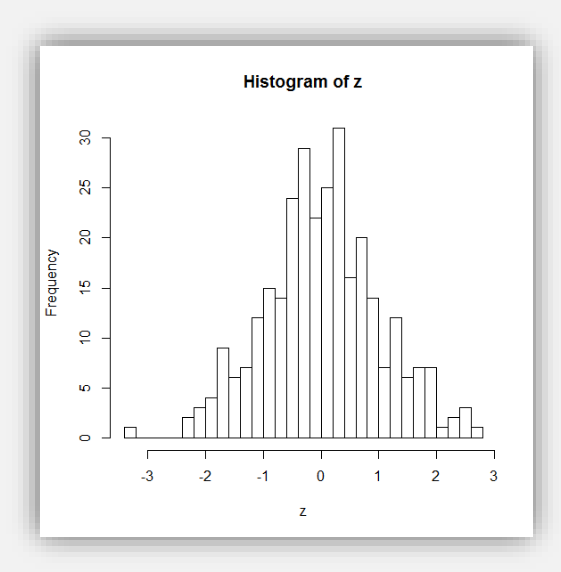
وهو طريقة وصفية أقل دقة من الطريقتين السابقتين، من الممكن الاستئناس بها للتأكد من طبيعية البيانات، وتعتمد هذه الطريقة على رسم مخطط انتشار للبيانات بعد ترتيبها تصاعدياً مع قيمها المعيارية، فإذا كان للانتشار شكل خطي حول مستقيم الطبيعية تكون البيانات طبيعية، وكلما ابتعدت البيانات عن مستقيم الطبيعية دلنا هذا على عدم التزام البيانات بالتوزيع الطبيعي.

يتم رسم مخطط QQ في R بالشكل:

التعليمة	
qqnorm(x) qqline(x)	
الناتج	
مخطط QQ يدل على أن العينة لدينا تتوزع وفق التوزيع الطبيعي	مخطط QQ يدل على أن العينة لدينا لا تتوزع وفق التوزيع الطبيعي

المدرج التكراري (Histogram):

سبق الكلام عنه في الفصل الخامس، وهو طريقة أقل جودة من أول طريقتين، وحتى تلتزم البيانات بالتوزيع الطبيعي يجب أن يكون لهذا المخطط شكلاً يشابه الشكل الآتي:



اختبارات تجانس التباينات (Homogeneity of Variance):

توجد العديد من اختبارات تجانس التباينات وأشهرها اختبار بارتليت Bartlett واختبار ليفين Levene:

اختبار بارتليت (Bartlett's Test):

يستخدم عندما تكون البيانات تتوزع وفق التوزيع الطبيعي ومقسمة لعدة مجموعات حسب عامل محدد ونريد اختبار الفرضية:

$$H_0: \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_r^2$$

مقابل:

$$H_1: \sigma_i^2 \neq \sigma_j^2 \ for\ some\ i, j$$

فإذا كانت لدينا Data frame اسمها myData فيها المتغير المدروس Y والعامل X يمكن تطبيق الاختبار السابق باستخدام R بالشكل:

```
bartlett.test(Y ~ X, data=myData)
```

فإذا كانت P>0.05 فإن البيانات متجانسة.

<u>مثال:</u> ولد 200 قيمة عشوائياً وقسمها إلى مجموعتين a و b وطبق اختبار بارتليت للتجانس.

مثال

```
y<-rnorm(200)
x<-c(rep(1,100),rep(2,100))
x<-factor(x,levels=c(1,2),labels=c("a","b"))
myData<-data.frame(y,x)
bartlett.test(y~x,myData)
```

الناتج

```
        Bartlett test of homogeneity of variances

data:  y by x
Bartlett's K-squared = 0.10146, df = 1, p-value = 0.7501
```

التفسير:

نلاحظ أن p>0.05 وبالتالي إن البيانات A والبيانات B متجانسة.

اختبار ليفين (Levene's Test):

يجب تحميل الحزمة car قبل تطبيق اختبار ليفين والذي يستخدم عندما تكون البيانات لا تتوزع وفق التوزيع الطبيعي ومقسمة لعدة مجموعات حسب عامل محدد ونريد اختبار الفرضية:

$$H_0: \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_r^2$$

مقابل:

$$H_1: \sigma_i^2 \neq \sigma_j^2 \ for\ some\ i, j$$

فإذا كانت لدينا Data frame اسمها myData فيها المتغير المدروس Y والعامل X يمكن تطبيق الاختبار السابق باستخدام R بالشكل:

```
library(car)
leveneTest (Y ~ X, data=myData)
```

فإذا كانت P>0.05 فإن البيانات متجانسة.

مثال: ولد 200 قيمة عشوائياً وقسمها إلى مجموعتين a و b وطبق اختبار ليفين للتجانس.

مثال

```
y<-rnorm(200)
x<-c(rep(1,100),rep(2,100))
x<-factor(x,levels=c(1,2),labels=c("a","b"))
myData<-data.frame(y,x)
leveneTest(y~x,myData)
```

الناتج

```
Levene's Test for Homogeneity of Variance (center = median)
        Df   F value   Pr(>F)
group   1    0.1099    0.7406
      198
```

التفسير:

نلاحظ أن p>0.05 وبالتالي إن البيانات A والبيانات B متجانسة.

الفصل الثامن
مستويات القياس
(Scales of Measurement)

بعد جمع الباحث للبيانات التي سيجري عليها التحليل الإحصائي واختبار الفرضيات يجب عليه معرفة نوع البيانات المستخدمة، حيث أنه لكل نوع من البيانات اختباراته الخاصة وعدم معرفة الباحث بذلك يجعله يقع في أخطاء جسيمة في بحثه وبالتالي تكون نتائج البحث مضللة.

يمكن تصنيف البيانات إلى نوعين رئيسيين هما البيانات النوعية Qualitative Data والبيانات الكمية Quantitative Data.

البيانات النوعية (Qualitative Data):

وتنقسم إلى نوعين هما البيانات الاسمية Nominal Data والبيانات الرتبية Ordinal Data ، أما البيانات الاسمية فهي تكون في صورة غير عددية، أي لا يمكن قياسها، وتتكون من فئات متشابهة تحمل نفس الخصائص لا يتم التفاضل بينها مثل الجنس الذي يتكون من فئتين، الذكور ونرمز لهم بالرقم (١) والإناث ونرمز لهن بالرقم (٢)، أو مثلاً السؤال الذي تكون إجابته "نعم " ونرمز له بالرقم (١) و" لا" ونرمز له بالرقم (٥)، في هذا النوع من البيانات لا نقوم سوى بحساب التكرارات والتمثيل البياني فقط، والخطأ الذي يقع فيه الباحث هو إجراء عمليات حسابية على البيانات الاسمية.

أما البيانات الترتيبية فهي أيضا تكون في صورة غير عددية ولا يمكن إجراء العمليات الحسابية عليها والفرق بينها وبين البيانات الاسمية هي عملية المفاضلة والترتيب بين طبقات المتغير مثل المستوى التعليمي (ابتدائي (١)، إعدادي (٢)، ثانوي (3)، جامعي فأكثر (4)) وفي هذه الحالة قد نستطيع تفسير المتوسط الحسابي أو الوسيط وبعض المقاييس الأخرى، إلا أن أي عملية حسابية على تلك البيانات ليس لها معنى.

البيانات الرقمية أو الكمية (Quantitative Dat):

تنقسم أيضا إلى نوعين هما البيانات الفتروية Interval Data والبيانات النسبية Rational Data ، أما البيانات الفتروية فتكون في صورة عددية ويمكن إجراء العمليات الحسابية عليها مثل المتوسط الحسابي والانحراف المعياري وغيرها ويمتاز هذا المقياس بتساوي المسافات بين الرتب حيث أنه يسمي أحيانا " بمقياس المسافة " ، ويستخدم هذا المقياس كثيراً في العلوم التربوية والنفسية والاجتماعية مثل قياس الذكاء وغيرها ، والجدير بالذكر أن هذا المقياس لا يعني الصفر فيه عدم وجود الخاصية فدرجة طالب تساوي صفر مثلا لا يعني أنه لا يعرف شيئاً في المقرر، كما أن درجة الحرارة صفر لا تعني انعدام ظاهرة الحرارة.

وأخيراً البيانات النسبية وهي أعلى مستوى من أنواع البيانات السابقة حيث يمتاز المستوى النسبي بكافة صفات المستويات السابقة بالإضافة لخاصية النسبية والتي تعني إن للصفر خاصية العدم أي خاصية انعدام الظاهرة، مثل سرعة السيارة التي تساوي صفر تعني أن السيارة متوقفة ، أو أن وزن شخص يساوي 60 كيلو جرام هو ضعف وزن شخص وزنه 30 كيلو جرام.

وبهذا نلخص هذا الفص حسب النظرية التي قام بتطويرها عالم علم النفس ستانلي سميث ستيفنس (Stanley Smith Stevens) و هي نظرية أنواع القياس 1946، حيث قال ستيفنس بأن القياس في العلوم يأخذ أحد الأشكال الآتية:

- الاسمي Nominal.
- الرتبي Ordinal.
- الفتروي Interval.
- النسبي Ratio.

الفصل التاسع
مقارنة المجموعات
(Comparing Groups)

اختبار ستيودينت للعينة الواحدة (One Sample t-test):

يستخدم اختبار ستيودينت للعينة الواحدة لاختبار الفرضية:

$$H_0: \bar{x} = m$$

مقابل:

$$H_1: \bar{x} \neq m$$

أي لاختبار اختلاف متوسط العينة $\bar{x}$ عن قيمة ما مثل m، ومن تطبيقاته في الحياة العملية اختبار القيمة m فيما إذا كانت شاذة أو لا.

للتطبيق باستخدام R نكتب التعليمة:

```
t.test(x,mu=m)
```

شروط تطبيق اختبار t لعينة واحدة
1. أن يكون متغير الدراسة كمياً.
2. أن تتوزع البيانات وفق التوزيع الطبيعي.
3. عدم وجود قيمة شاذة.
4. أن يكون حجم العينة يشكل 5% على الأقل من حجم المجتمع.

مثال:

لتكن لدينا قياسات سكر الدم لمجموعة من المرضى الذين تم علاجهم باستخدام الدواءين Lantus+R، والمطلوب معرفة فيما إذا كان سكر الدم للعينة منضبطاً، علماً أن متوسط سكر الدم الطبيعي لمرضى السكر يعتبر 110.

Lantus+R	140	90	110	125	111	128	113	89	110

الحل
x<-c(110,89,113,128,111,125,110,90,140) t.test(x,mu=110)
الناتج
One Sample t-test data: x 1. t = 0.5197, df = 8, p-value = 0.6174 2. alternative hypothesis: true mean is not equal to 110 3. 95 percent confidence interval: 4. 125.7077 100.0701 5. sample estimates: 6. mean of x 7. 112.88895

التفسير:

في السطر الأول يعرض R إحصائية الاختبار t وعدد درجات الحرية و p-value (المعنوية) وقد كانت قيمة المعنوية p-value=0.6174 والتي هي أكبر من 0.05 وبالتالي لا نستطيع أن نرفض الفرضية الابتدائية التي تقول بعدم وجود اختلاف معنوي لسكر الدم عند المرضى عن سكر الدم الطبيعي وذلك عند مستوى المعنوية 5%، السطر الثاني يخبرنا بأن الفرضية البديلة تنص على أن متوسط سكر الدم للمجموعة يختلف عن 110، ويقصد بالسطر الثالث أنه سيتم إظهار 95% فترة ثقة لمتوسط سكر الدم للمرضى، وبالسطر الرابع تم عرض فترة الثقة والتي كانت [100.07, 125.71] والتي تعني أنه وبثقة 95% في المجتمع الأصلي الذي سحبنا منه عينتنا سيكون متوسط سكر الدم بين 100.07 و 125.71، أما السطر الخامس فمعناه: مقدرات العينة، والسطر السادس يعني متوسط x والذي تم إظهاره في السطر السابع، ويمكن تقدير متوسط المجتمع الذي سحبت منه العينة x بـ 112.89.

ثانياً: اختبار ستيودينت للعينتين المستقلتين (Independent Samples t-test):

يستخدم اختبار ستيودينت للعينتين المستقلتين لاختبار الفرضية:

$$H_0: \bar{x} = \bar{y}$$

مقابل:

$$H_1: \bar{x} \neq \bar{y}$$

أي لاختبار اختلاف متوسط العينة $\bar{x}$ عن متوسط العينة $\bar{y}$، أي لمعرفة إذا كان الفرق بين $\bar{x},\bar{y}$ هو فرق ذو أهمية إحصائية أو أنه فقط بمجرد الصدفة.

هناك حالة أعم لاختبار ستيودينت للعينتين المستقلتين وتصاغ فرضيتاه بالشكل:

$$H_0: \bar{x} - \bar{y} = m$$

مقابل:

$$H_1: \bar{x} - \bar{y} \neq m$$

والذي يدرس معنوية كون الفرق بين $\bar{x},\bar{y}$ يساوي مقدار ثابت مثل m.

للتطبيق باستخدام R نكتب التعليمة:

```
t.test(x,y,mu=m)
```

ويمكن إهمال الوسيط الثالث في حال كان m=0.

شروط تطبيق اختبار t لعينتين مستقلتين
1. أن يكون متغير الدراسة كمياً.   2. أن تتوزع البيانات لكل من العينتين وفق التوزيع الطبيعي.   3. تجانس تباين العينتين.   4. أن تكون العينتان مستقلتين.   5. عدم وجود قيمة شاذة.   6. أن يكون حجم العينة يشكل 5% على الأقل من حجم المجتمع.

<u>مثال:</u>

لتكن لدينا قياسات سكر الدم لمجموعتين من المرضى حيث تأخذ المجموعة الأولى الدواء Lantus وتأخذ المجموعة الثانية الدواءين Lantus+R:

140	149	164	138	121	111	139	128	111	Lantus
110	89	113	128	111	125	110	90	140	Lantus+R

والمطلوب معرفة أي الدواءين هو الأفضل.

الحل

```
x<-c(140,149,164,138,121,111,139,128,111)
y<-c(110,89,113,128,111,125,110,90,140)
t.test(x,y)
```

الناتج
Welch Two Sample t-test **data: x and y** **t = 2.5477, df = 15.959, p-value = 0.02153 .1** **alternative hypothesis: true difference in means is not equal to 0 .2** **95 percent confidence interval: .3** **37.663323 3.447788 .4** **sample estimates: .5** **mean of y mean of x .6** **112.8889 133.4444 .1**

التفسير:

في السطر الأول يعرض R إحصائية الاختبار t وعدد درجات الحرية و p-value (المعنوية) وقد كانت قيمة p-value=0.02153 والتي هي أقل من 0.05 وبالتالي نرفض الفرضية الابتدائية التي تقول بعدم وجود اختلاف معنوي بين سكر الدم في كل من المجموعتين؛ أي؛ يوجد اختلاف معنوي بين سكر الدم باستخدام العلاج الأول وسكر الدم باستخدام العلاج الثاني وذلك عند مستوى المعنوية 5%، السطر الثاني يخبرنا بأن الفرضية البديلة تنص على أن الفرق بين متوسطي المجموعتين يختلف إحصائياً عن الصفر، ويقصد بالسطر الثالث أنه سيتم إظهار 95% فترة ثقة للفرق بين متوسطي المجموعتين، وبالسطر الرابع تم عرض فترة الثقة والتي كانت [3.45,37.66] والتي تعني أنه وبثقة 95% في المجتمع الأصلي الذي سحبنا منه عينتنا لن يكون الفرق بين متوسطي المجموعتين أقل من 3.45 ولا أكبر من 37.66، أما السطر الخامس فمعناه: مقدرات العينة، والسطر السادس يعني متوسط x ومتوسط y والتي تم إظهار كل منها في السطر السابع، ويمكن تقدير متوسط المجتمع الذي سحبت منه العينة x بـ 133.44 ومتوسط المجتمع الذي سحبت منه العينة y بـ 112.89.

ثالثاً: اختبار ستيودينت للعينة المزدوجة (Paired- Sample t-test):

يستخدم اختبار ستيودينت للعينة المزدوجة (العينتين المرتبطتين) لاختبار الفرضية:

$$H_0: \bar{x}_1 = \bar{x}_2$$

مقابل:

$$H_1: \bar{x}_1 \neq \bar{x}_2$$

أي لاختبار اختلاف متوسط العينة $\bar{x}_1$ في ظرف ما عن متوسط العينة نفسها في ظرف آخر $\bar{x}_2$، ومعرفة إذا كان هذا الفرق هو فرق ذو أهمية إحصائية أو أنه فقط بمجرد الصدفة.

هناك حالة أعم لاختبار ستيودينت للعينة المزدوجة وتصاغ فرضيتاه بالشكل:

$$H_0: \bar{x}_1 - \bar{x}_2 = m$$

مقابل:

$$H_1: \bar{x}_1 - \bar{x}_2 \neq m$$

والذي يدرس معنوية كون الفرق بين $\bar{x}_1, \bar{x}_2$ يساوي مقدار ثابت مثل m.

للتطبيق باستخدام R نكتب التعليمة:

```
t.test(x,y,mu=m,paired=T)
```

ويمكن إهمال الوسيط الثالث في حال كان m=0.

شروط تطبيق اختبار t للعينة المزدوجة
1. أن يكون متغير الدراسة كمياً. 2. أن يتوزع فرق العينتين وفق التوزيع الطبيعي. 3. أن تكون العينتان عبارة عن عينة واحدة مقاسة في ظرفين مختلفين. 4. عدم وجود قيمة شاذة. 5. أن يكون حجم العينة يشكل 5% على الأقل من حجم المجتمع.

<u>مثال:</u>

قمنا بتطبيق نظام حمية على عينة من النساء لمدة شهر وقمنا بتسجيل أوزانهن قبل وبعد الحمية فكانت النتائج كما يلي:

قبل	65	66	62	59	62	74	63	69	65
بعد	60	60	59	58	54	67	58	62	60

والمطلوب معرفة فيما إذا كانت الحمية مجدية.

الحل
x<-c(65,69,63,74,62,59,62,66,65) y<-c(60,62,58,67,54,58,59,60,60) t.test(x,y,paired=T)
الناتج
Paired t-test data: x and y t = 7.2308, df = 8, p-value = 8.972e-05 .1 alternative hypothesis: true difference in means is not equal to 0 .2 95 percent confidence interval: .3 3.556775 6.887670 .4 sample estimates: .5 mean of the differences .6 5.222222 .2

التفسير:

في السطر الأول يعرض R إحصائية الاختبار t وعدد درجات الحرية و p-value (المعنوية) وقد كانت قيمة p-value=0.0000897 والتي هي أقل من 0.05 وبالتالي نرفض الفرضية الابتدائية التي تقول بعدم وجود اختلاف معنوي بين الأوزان في كلتي الحالتين.

أي يوجد اختلاف معنوي بين أوزان النساء قبل الحمية وأوزان النساء بعد الحمية وذلك عند مستوى المعنوية 5%، السطر الثاني يخبرنا بأن الفرضية البديلة تنص على أن الفرق بين متوسط المجموعتين يختلف إحصائياً عن الصفر، ويقصد بالسطر الثالث أنه سيتم إظهار 95% فترة ثقة للفرق بين متوسطي المجموعتين، وبالسطر الرابع تم عرض فترة الثقة والتي كانت [3.56,6.89] والتي تعني أنه وبثقة 95% في المجتمع الأصلي الذي سحبنا منه عينتنا لن يكون الفرق بين متوسطي المجموعتين أقل من 3.56 ولا أكبر من 6.89، أما السطر الخامس فمعناه: مقدرات العينة، والسطر السادس يعني متوسط الفرق بين المجموعتين والذي سيتم إظهاره في السطر السابع، ويمكن تقدير متوسط الفرق بين أوزان النساء قبل وبعد الحمية بـ 5.22.

تحليل التباين أحادي الاتجاه (One Way ANOVA):

يستخدم اختبار تحليل التباين أحادي الاتجاه لدراسة وجود فروق معنوية بين عدة مجموعات، أي لاختبار الفرضية:

$$H_0: \bar{x}_1 = \bar{x}_2 = \cdots = \bar{x}_r$$

مقابل:

$$H_1: \bar{x}_i \neq \bar{x}_j \; for \; some \; i, j$$

ويمكن إجراء الاختبار باستخدام R بالخطوات الآتية:

1. نعرف data frame.
2. نقوم بعملية stack لـ data frame ونحفظها في متغير جديد وليكن xs.
3. نستخدم التعليمة:

```
aov(values~ind,data=xs)
```

شروط تطبيق اختبار One Way ANOVA
1. أن يكون متغير الدراسة كمياً. 2. أن تتوزع البيانات لكل من العينات وفق التوزيع الطبيعي. 3. تجانس تباين العينات 4. أن تكون العينات مستقلة. 5. عدم وجود قيمة شاذة. 6. أن يكون حجم العينة يشكل 5% على الأقل من حجم المجتمع.

مثال:

لدراسة الاختلاف في معدلات الطلاب باختلاف طريقة التدريس كانت لدينا النتائج الآتية:

A	B	C
75	84	88
77	87	82
72	83	87
78	77	89
89	79	83
79	82	85
81	79	89

حيث تمثل A طريقة التدريس التقليدية، و B التدريس مع جهاز إسقاط، و C التدريس مع مذاكرات دورية.

الحل

```
a<-c(75,77,72,78,89,79,81)
b<-c(84,87,83,77,79,82,79)
c<-c(88,82,87,89,83,85,89)
x<-data.frame(a,b,c)
xs<-stack(x)
anova<-aov(values~ind,data=xs)
summary(anova)
```

الناتج

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
ind	2	196.6	98.29	6.017	0.00997 **
Residuals	18	294.0	16.33		

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

والذي يهمنا من الناتج هو القيمة p=0.00997 والتي هي أقل من 0.05 وبالتالي نستنتج أن طريقتين على الأقل من الطرائق الثلاثة تختلفان عن بعضهما البعض اختلافاً معنوياً، وبالتالي نسأل السؤال الآتي: أي الطرائق هي التي تختلف عن بعضها البعض؟

اختبار Tukey HSD:

يستخدم اختبار Tukey بعد تطبيق اختبار ANOVA وملاحظة أن p<0.05 وذلك لمعرفة أي المجموعات هي التي تختلف عن بعضها البعض اختلافاً معنوياً، وذلك بالشكل:

```
TukeyHSD(anova)
```

حيث anova هو متحول حفظت فيه نتيجة اختبار ANOVA، وفي مثالنا السابق تكون النتيجة:

الحل

```
TukeyHSD(anova)
```

الناتج

Fit: aov(formula = values ~ ind, data = xs)

ind$

diff	lwr	upr	p adj	
b-a	2.857143	-2.656160	8.370445	0.4013131
c-a	7.428571	1.915269	12.941874	0.0078367
c-b	4.571429	-0.941874	10.084731	0.1145678

التفسير:

يمثل السطر b-a مقارنة المجموعتين A و B، فالقيمة 2.857 تمثل الفرق في المتوسطات B-A وتمثل القيمة 2.656- الحد الأدنى لـ95% فترة ثقة للفرق بين المتوسطين، وتمثل القيمة 8.370 الحد الأعلى لـ95% فترة ثقة للفرق بين المتوسطين، والقيمة الأخيرة 0.401 تمثل معنوية هذا الفرق وهو غير معنوي كونه أكبر من 0.05، ونلاحظ من الجدول السابق أن المجموعتين C و A هما المجموعتان الوحيدتان اللتان تختلفان معنوياً عن بعضهما البعض كون p=0.007<0.05.

تحليل التباين ثنائي الاتجاه (Two Way ANOVA):

يستخدم تحليل التباين ثنائي الاتجاه لدراسة تأثير عاملين على متغير محدد، مثل تأثير الجنس والحالة الاجتماعية على معدلات الطلاب أو تأثير طريقة الري ونوع السماد على كمية الإنتاج، ويمكن إجراء الاختبار باستخدام R بالخطوات الآتية:

1. نعرف data frame باسم ما وليكن myData تحتوي على المتغير التابع values والعامل الأول factor1 والعامل الثاني factor2.
2. نستخدم التعليمة:

```
aov(values~factor1*factor2,data=myData)
```

شروط تطبيق اختبار Two Way ANOVA
1. أن يكون متغير الدراسة كمياً.
2. أن تتوزع البيانات لكل من العينات وفق التوزيع الطبيعي.
3. تجانس تباين العينات.
4. عدم وجود قيمة شاذة.
5. أن يكون حجم العينة يشكل 5% على الأقل من حجم المجتمع.

مثال:

لنفرض أننا نود دراسة تأثير كل من الجنس وطريقة العلاج على سكر الدم، وكان لدينا:

Glucose	Gender	Medicine
168	Male	**Lantus+R**
187	female	**Mix**
165	Male	**Lantus**
198	female	Lantus+R
178	Male	Mix
128	female	Lantus
145	Male	Lantus+R
197	female	Mix
188	Male	Lantus
169	female	Lantus+R
168	Male	Mix
180	female	Lantus
210	Male	Lantus+R
211	female	Mix
114	Male	Lantus
112	female	Lantus+R
129	Male	Mix
90	female	Lantus
115	Male	Lantus+R
120	female	Mix
100	Male	Lantus

سنطبق تحليل التباين ثنائي الاتجاه لأنه لدينا عاملان ومتغير تابع:

الحل

```
Glucose<-
c(100,120,115,90,129,112,114,211,210,180,168,169,188,197,145,128,178,198,165,187,168)
Gender<-c(1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1)
Medicine<-c(1,1,1,1,1,1,1,2,2,2,2,2,2,2,3,3,3,3,3,3,3)
myData<-data.frame("Glucose"=Glucose,"Medicine"=Medicine,"Gender"=Gender)
myData$Medicine<-factor(myData $Medicine,levels=c(1,2,3),labels=c("Lantus +
R","Mix","Lantus"))
myData$Gender<-factor(myData $Gender,levels=c(1,2),labels=c("Male","Female"))
anova<- aov(Glucose~Medicine*Gender,myData)
summary(anova)
```

الناتج

```
                 Df   Sum Sq   Mean Sq   F value    Pr(>F)
Medicine         2    22376    11188     26.967     1.08e-05 ***
Gender           1    0        0         0          0.988
Medicine:Gender  2    173      86        0.208      0.815
Residuals        15   6223     415
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01
‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

من الجدول السابق نلاحظ أنه لنوع العلاج فقط تأثير معنوي على معدل سكر الدم كون p<0.05، وليس هناك اختلاف بمستويات سكر الدم باختلاف الجنس، ولا تفاعل مشترك بين الجنس ونوع العلاج، <u>لكن السؤال: ما هو العلاج الأكثر تأثيراً؟</u>

عودة لاختبار Tukey HSD:

نستطيع تطبيق اختبار TukeyHSD مع Two Way ANOVA كما طبقناه مع One Way ANOVA لكن يفضل عدم إدخال المتغيرات غير المعنوية، وذلك بالشكل:

```
TukeyHSD(anova,wich=”significantFactors”)
```

وفي مثالنا السابق يكون لدينا:

الحل
TukeyHSD(anova,which="Medicine")
الناتج

```
 Tukey multiple comparisons of means
   95% family-wise confidence level

Fit: aov(formula = Glucose ~ Medicine * Gender, data = myData)

$Medicine
                      diff        lwr        upr     p adj
Mix-Lantus + R    77.57143   49.29180  105.851059  0.0000099
Lantus-Lantus + R 55.57143   27.29180   83.851059  0.0003589
Lantus-Mix       -22.00000  -50.27963    6.279631  0.1413899
```

ونلاحظ وجود فرق معنوي بين مستخدمي Mix Insulin و Lantus+R حيث كان **سكر** الدم لمستخدمي Lantus+R أكثر انضباطاً، وكذلك يوجد فرق معنوي بين مستخدمي Lantus ومستخدمي Lantus+R وكان **سكر** الدم لمستخدمي Lantus+R أيضاً أفضل.

الفصل العاشر
العلاقة بين المتغيرات والانحدار
(Correlation and Regression)

معامل ارتباط بيرسون (Pearson Correlation Coefficient):

يستخدم معامل ارتباط بيرسون لدراسة وجود علاقة بين متغيرين كميين X و Y وشدة هذه العلاقة، حيث يقع معامل الارتباط ضمن المجال [-1,1] وكلما اقتربت قيمته المطلقة من الواحد دلنا هذا على أن العلاقة أقوى، أما عند اقتراب قيمته المطلقة من النصف تكون العلاقة متوسطة، وعند اقتراب قيمته من الصفر تكون العلاقة ضعيفة، والإشارة الموجبة لمعامل الارتباط تدل على أن العلاقة طردية، أما الإشارة السالبة فتدل على أن العلاقة عكسية، بعد حساب معامل الارتباط R يتم اختبار الفرضية:

$$H_0: R = 0$$

مقابل

$$H_1: R \neq 0$$

ونقوم بدراسة العلاقة بين متغيرين X,Y وفق معامل ارتباط بيرسون باستخدام التعليمة:

```
cor.test(x,y)
```

شروط تطبيق معامل ارتباط بيرسون الخطي
1. أن يكون متغيرا الدراسة كميين. 2. أن تتوزع البيانات لكل من العينتين وفق التوزيع الطبيعي. 3. أن تكون العلاقة بين المتغيرين خطية (نتحقق من ذلك برسم مخطط الانتشار). 4. ثبات التباين Homoscedasticity حول خط الانتشار (يجب أن يكون الانتشار على شكل سيجار تقريباً). 5. عدم وجود قيمة شاذة. 6. أن يكون حجم العينة يشكل 5% على الأقل من حجم المجتمع.

مثال:

لنولد شعاعين X,Y عشوائياً وندرس العلاقة بينهما:

الحل
x<-rnorm(50) y<-rnorm(50) cor.test(x,y)
الناتج
Pearson's product-moment correlation data: x and y t = 0.1508, df = 48, p-value = 0.8808 alternative hypothesis: true correlation is not equal to 0 95 percent confidence interval: -0.2581507 0.2983015 sample estimates: cor 0.02176061

والذي يهمنا أولاً هو القيمة p=0.8808 والتي هي أكبر من 0.05 وبالتالي نستنتج أنه لا توجد علاقة معنوية بين X و Y، كما يعطينا R فترة ثقة 95% لمعامل الارتباط وهي [-0.258,0.298] أي أنه وبثقة 95% سيكون R للمجتمع بين -25.8% و 29.8% حيث كان R من العينة 0.022 وهو صغير جداً وهو سبب كونه ليس معنوياً.

الارتباط لا يعني السببية (Correlation Versus Causality):

الارتباط يعني وجود علاقة بين المتغيرين X,Y لكن ليس من الضرورة أن هذه العلاقة تعني السببية، أي أن وجود علاقة بين X,Y لا تعني أن X يسبب Y أو أن Y يسبب X، بل قد يكون هناك متغير ما مثل T هو الذي يسببهما.

من الأمثلة التي طرحت هذا النقاش أن أحد الباحثين وجد علاقة بين تناول المثلجات والقيام بالجرائم في New York، **وطرح السؤال الآتي**: **هل تناول المثلجات يجعل الشخص مجرماً؟**

الإجابة عن هذا السؤال كانت لا، إنما المتغيران السابقان متأثران بالطقس، ففي الهجمات الحرارية الشديدة يرتفع كل من معدل تناول المثلجات ومعدل الجرائم مما يشكل ارتباطاً زائفاً بفعل متغير وسيط وهو الطقس، ويمكن استبعاد أثر هذا المتغير باستخدام ما يعرف بالارتباط الجزئي.

الارتباط الجزئي (Partial Correlation):

كما سبق وقلنا إن الارتباط الجزئي يستخدم لدراسة العلاقة بين متغيرين X و Y بعد إزالة أثر متغير آخر Z، ولتطبيق الارتباط الجزئي في R سنحتاج للحزمة ppcor، والتعليمة اللازمة لحساب الارتباط الجزئي هي كالآتي:

```
install.packages("ppcor")
library("ppcor")
pcor.test(x,y,z)
```

<u>مثال:</u>

ولد ثلاثة أشعة X,Y,Z واحسب العلاقة بين X,Y مع استبعاد أثر Z:

الحل

```
x<-rnorm(50)
y<-rnorm(50)
z<-rnorm(50)
install.packages("ppcor")
library("ppcor")
pcor.test(x,y,z)
```

الناتج

	estimate	p.value	statistic	n	gp	Method
1	0.2576389	0.07390174	1.827994	50	1	pearson

ونلاحظ مما سبق أنه لا علاقة معنوية بين X و Y ($p>0.05$) بعد استبعاد أثر Z.

<u>ملاحظة:</u>

يمكن إيجاد مصفوفة الارتباط لأكثر من شعاعين باستخدام نفس التعليمة cor(A) حيث أن A هو متغير معرف على أنه matrix كما يمكن إيجاد مصفوفة التغاير باستخدام التعليمة cov(A).

الانحدار الخطي البسيط (Simple Linear Regression):

يقصد بالانحدار الخطي البسيط دراسة تأثير المتغير X والذي يدعى المتغير المستقل (Independent Variable) على المتغير Y والذي يدعى المتغير التابع (Dependent Variable).

يعطى نموذج الانحدار الخطي بالشكل:

$$Y = \beta_0 + \beta_1 X + \epsilon \, ; \epsilon \sim N(0, \sigma^2)$$

حيث ندعو ϵ بالراسب أو الخطأ، ويهدف تحليل الانحدار إلى إيجاد مقدرات لكل من β_0, β_1 والتي تجعل مجموع مربعات الرواسب أصغر ما يمكن.

لتطبيق الانحدار الخطي البسيط باستخدام R نكتب التعليمة:

```
lm(y~x)
```

شروط تطبيق الانحدار الخطي البسيط
1. أن يكون كل من المتغير التابع والمتغير المستقل كميين. 2. عدم وجود علاقة غير خطية بين الرواسب و Y المقدرة. 3. استقلال الرواسب. 4. التوزيع الطبيعي للرواسب. 5. تجانس التباين. 6. عدم وجود قيم شاذة. 7. أن يكون حجم العينة كبيراً.

مثال:

أوجد معادلة الانحدار الخطي البسيط التي تمثل دور الطول بالتنبؤ بالوزن بالاعتماد على العينة الآتية:

الطول	163	168	169	174	175	170	167	160
الوزن	61	67	65	78	74	73	65	58

الحل:

```
x<-c(160,167,170,175,174,169,168,163)
y<-c(58,65,73,74,78,65,67,61)
lm(y~x)
```

الناتج:

```
Call:
lm(formula = y ~ x)


Coefficients:
(Intercept)          x
  -145.851       1.269
```

التفسير:

يهمنا من الجدول السابق قيما intercept و x حيث تمثل intercept المعامل β_0 وتمثل قيمة x المعامل β_1 وبالتالي إن معادلة الانحدار التي تمثل إسهام الطول X بالتنبؤ بالوزن y هي:

$$Y = -145.851 + 1.269\,X$$

إن هذه المعادلة غير كافية إحصائياً، فيجب دراسة مدى كفاءة هذه المعادلة بالتنبؤ ومدى جودتها ومعنويتها، وهذا يتم بالشكل الآتي:

التعليمة:

```
reg<-lm(y~x)
summary(reg)
```

الناتج:

```
Call:
lm(formula = y ~ x)

1.  Residuals:
2.  Min          1Q        Median       3Q          Max
3.  -3.5766    -1.3266    -0.1358     1.4018     3.1546

4.  Coefficients:
                    i.  Estimate      Std. Error      t          value Pr(>|t|)
5.  (Intercept)   -145.8510       31.9907   -4.559        0.003854 **
6.  x                  1.2688         0.1901     6.676        0.000547 ***
7.  ---
8.  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

9.  Residual standard error: 2.546 on 6 degrees of freedom
10. Multiple R-squared:  0.8813,    Adjusted R-squared:  0.8616
11. F-statistic: 44.57 on 1 and 6 DF,  p-value: 0.0005471
```

التفسير:

السطر الأول يوضح أن المخرجات الآتية هي للبواقي (Residuals)، أي الفرق بين القيم الفعلية والقيم المتنبأة، فيظهر في السطر الثالث على الترتيب أصغر راسب، والربيع الأول للرواسب ووسيط الرواسب، والربيع الثالث للرواسب، وأكبر راسب.

في السطر الرابع يوضح أن المخرجات الآتية هي معاملات النموذج ويمثل السطر الخامس الحد الثابت من نموذج الانحدار وقيمته في مثالنا $\beta_0 = -145.851$ بخطأ معياري 31.99 وإحصاء اختبار معنوية هذا المعامل يخضع لتوزيع ستيودينت وقيمته $t = -4.559$ وهو معنوي كون $p - value < 0.05$.

يمثل السطر السادس المعامل β_1 الذي كانت قيمته $\beta_1 = 1.2688$ بخطأ معياري 0.1901 وإحصاء اختبار معنوية هذا المعامل يخضع أيضاً لتوزيع ستيودينت وقيمته $t = 6.676$ وهو معنوي كون $p - value < 0.05$.

في السطر الثامن يبين R لنا تفسيره لمعنوية المعاملات حيث يصطلح R الرمز *** للمعاملات ذات المعنوية العالية جداً والرمز ** للمعنوية العالية والرمز * للمعنوية العادية.

السطر التاسع يبين الخطأ المعياري للرواسب والذي قد كان 2.546 بـ 6 درجات حرية أما السطر العاشر فيبين قيمة معامل التحديد $R^2 = 0.8813$ والذي يعني بمثالنا أن 88.13% من التغير في الوزن هو بسبب التغير في الطول، أو أن الطول يفسر 88.13% من التغير في الوزن، كما يبين في السطر نفسه قيمة معامل التحديد المعدل والذي بلغ في نموذجنا 86.16% ويستخدم معامل التحديد المعدل بدلاً من معامل التحديد العادي في حال كان حجم العينة صغيراً.

السطر الأخير يظهر الإحصائية العامة عن معنوية النموذج حيث بلغت قيمة إحصاء الاختبار $F = 44.57$ بدرجة حرية واحدة للبسط وست درجات حرية للمقام وبمعنوية $p - value < 0.05$ أي أن النموذج المقترح هو نموذج معنوي.

في النهاية نستطيع كتابة معادلة الانحدار الممثلة لمثالنا بالشكل:

$$Weight = -145.851 + 1.269 * Height$$

الانحدار الخطي المتعدد (Multiple Linear Regression):

يقصد بالانحدار الخطي المتعدد دراسة تأثير (علاقة) عدة متغيرات $X_1, X_2, \dots, X_p$ تدعى المتغيرات المستقلة (Independent Variables) على متغير Y يدعى المتغير التابع (Dependent Variable).

يعطى نموذج الانحدار الخطي المتعدد بالشكل:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon \, ; \epsilon \sim N(0, \sigma^2)$$

لتطبيق الانحدار المتعدد باستخدام R نكتب التعليمة:

```
lm(y~x1+x2+...+xp)
```

شروط تطبيق الانحدار الخطي البسيط
1. أن يكون كل من المتغير التابع والمتغيرات المستقلة كمية.
2. عدم وجود علاقة غير خطية بين الرواسب و Y المقدرة.
3. استقلال الرواسب.
4. التوزيع الطبيعي للرواسب.
5. ثبات التباين.
6. عدم وجود مصاحبة خطية متعددة (Multicollinearity) بين المتغيرات المستقلة.
7. عدم وجود قيم شاذة.
8. أن يكون حجم العينة كبيراً.

<u>مثال:</u>

أوجد معادلة الانحدار الخطي البسيط التي تمثل دور الطول وسكر الدم بالتنبؤ بالوزن بالاعتماد على العينة الآتية:

الطول	170	163	167	175	162	178	172
الوزن	90	59	60	75	60	80	70
سكر الدم	165	94	89	119	100	128	113

الحل:

```
y<-c(70,80,60,75,60,59,90)
x1<-c(172,178,162,175,167,163,170)
x2<-c(113,128,100,119,89,94,165)
lm(y~x1+x2)
```

الناتج:

```
Call:
lm(formula = y ~ x1 + x2)


Coefficients:
(Intercept)          x1           x2  
   -65.4312      0.5425       0.3812
```

التفسير:

يهمنا من الجدول السابق قيم intercept و x1 وx2 حيث تمثل intercept المعامل β_0 وتمثل قيمة x1 المعامل β_1 وقيمة x2 المعامل β_2 وبالتالي إن معادلة الانحدار التي تمثل إسهام الطول X_1 وسكر الدم X_2 بالتنبؤ بالوزن Y هي:

$$Y = -65.43 + 0.54X_1 + 0.38X_2$$

إن هذه المعادلة غير كافية إحصائياً، فيجب دراسة مدى كفاءة هذه المعادلة بالتنبؤ ومدى جودتها ومعنويتها، وهذا يتم بالشكل الآتي:

الحل:

```
reg<-lm(y~x1+x2)
summary(reg)
```

الخرج:

```
Call:
lm(formula = y ~ x1 + x2)

1. Residuals:
2.        1         2         3         4         5         6         7 
3. -0.96309  0.06308 -0.58168  0.12187  0.89944  0.16328  0.29711 

4. Coefficients:
              Estimate   Std. Error    t value     Pr(>|t|)    
5. (Intercept) -65.43125    9.28691     -7.046     0.002139 ** 
6. x1            0.54252    0.05879      9.228     0.000766 ***
7. x2            0.38125    0.01354     28.160     9.46e-06 ***
8. ---
9. Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

10. Residual standard error: 0.7431 on 4 degrees of freedom
11. Multiple R-squared:  0.9974,    Adjusted R-squared:  0.9961 
12. F-statistic:   762 on 2 and 4 DF,  p-value: 6.853e-06
```

التفسير:

السطر الأول يوضح أن المخرجات الآتية هي للبواقي (الرواسب) (Residuals)، أي الفرق بين القيم الفعلية والقيم المتنبأة، فيظهر في السطر الثالث قيم الرواسب.

في السطر الرابع يوضح أن المخرجات الآتية هي معاملات النموذج ويظهر في كل من السطر الخامس والسادس والسابع قيم المعاملات وخطأها المعياري وإحصائية اختبارها ومعنويتها على الترتيب، ونلاحظ في مثالنا أن كافة المعاملات معنوية.

في السطر التاسع يبين R لنا تفسيره لمعنوية المعاملات حيث يصطلح R الرمز *** للمعاملات ذات المعنوية العالية جداً والرمز ** للمعنوية العالية والرمز * للمعنوية العادية.

السطر العاشر يبين الخطأ المعياري للرواسب والذي قد كان 0.7431 بـ 4 درجات حرية أما السطر الحادي عشر فيبين قيمة معامل التحديد $R^2 = 0.9974$ والذي يعني بمثالنا أن 99.74% من التغير في الوزن هو بسبب التغير في الطول وسكر الدم، أو أن الطول وسكر الدم يفسران 99.74% من التغير في الوزن، كما يبين في السطر نفسه قيمة معامل التحديد المعدل والذي بلغ في نموذجنا 99.61% ويستخدم معامل التحديد المعدل بدلاً من معامل التحديد العادي في حال كان حجم العينة صغيراً.

السطر الأخير يظهر الإحصائية العامة عن معنوية النموذج حيث بلغت قيمة إحصاء الاختبار $F = 762$ بدرجتي حرية للبسط وأربع درجات حرية للمقام وبمعنوية $p < 0.05$ أي أن النموذج المقترح هو نموذج معنوي.

التحقق من شروط الانحدار (Checking Regression Assumptions):

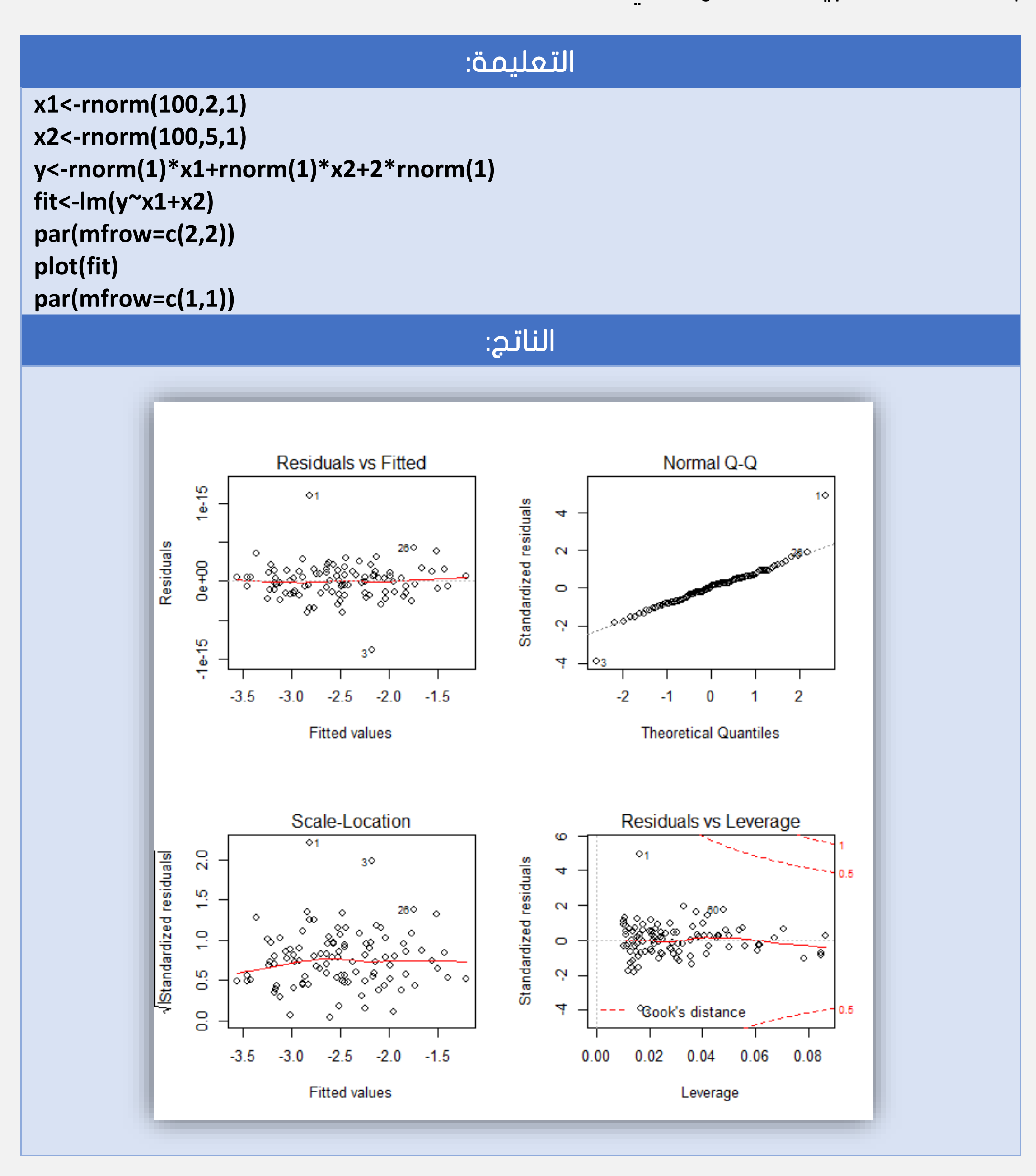
إن تحليل الانحدار حساس كثيراً لفرضيات تطبيقه النظرية، خاصة إذا كان الهدف استخدامه في التنبؤ، فلا يمكننا الاكتفاء بأن نجد أن معادلة الانحدار معنوية لاستخدامها في التنبؤ دون التأكد من تحقق شروط تطبيقه.

ستحتاج للحزمة car في هذه الفقرة فلا تنس تحميلها، وسنقوم بكافة الاختبارات بالاعتماد على بيانات المثال الآتي:

التعليمة:

```
x1<-rnorm(100,2,1)
x2<-rnorm(100,5,1)
y<-rnorm(1)*x1+rnorm(1)*x2+2*rnorm(1)
fit<-lm(y~x1+x2)
par(mfrow=c(2,2))
plot(fit)
par(mfrow=c(1,1))
```

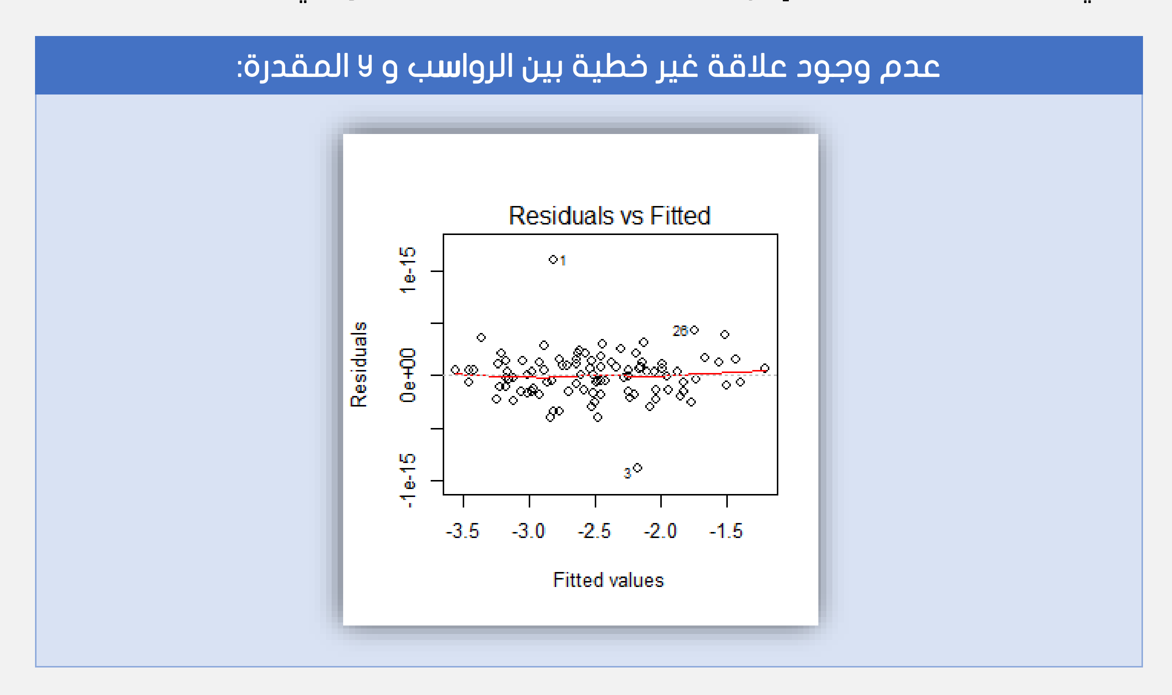
الناتج:



الرسم السابق يظهر لنا الرواسب في أربع طرائق مختلفة، وسنبين فائدة كل منها:

عدم وجود علاقة غير خطية بين الرواسب وال y المقدرة (Nonlinearity):

الذي يختبر هذه الفرضية هو الرسم المسمى Residuals vs fitted يجب ألا تكون هناك علاقة غير خطية بين الرواسب وال y المقدرة لأن هذا يعني أنه هناك علاقة غير خطية بين المتغيرات المستقلة والمتغير التابع ولم يستطع النموذج تفسيرها، مما أدى لظهورها في الرواسب، فإذا رأيت أن الرواسب تنتشر بشكل عشوائي فإن هذا مؤشر جيد على التزام نموذجك بهذه الفرضية، إما إذا كان للرواسب انتشار على شكل قطع مكافئ أو شكل أسي مثلاً، فعليك أن تتخذ إجراء مناسباً لحل هذه المشكلة، وفي مثالنا كان لدينا:

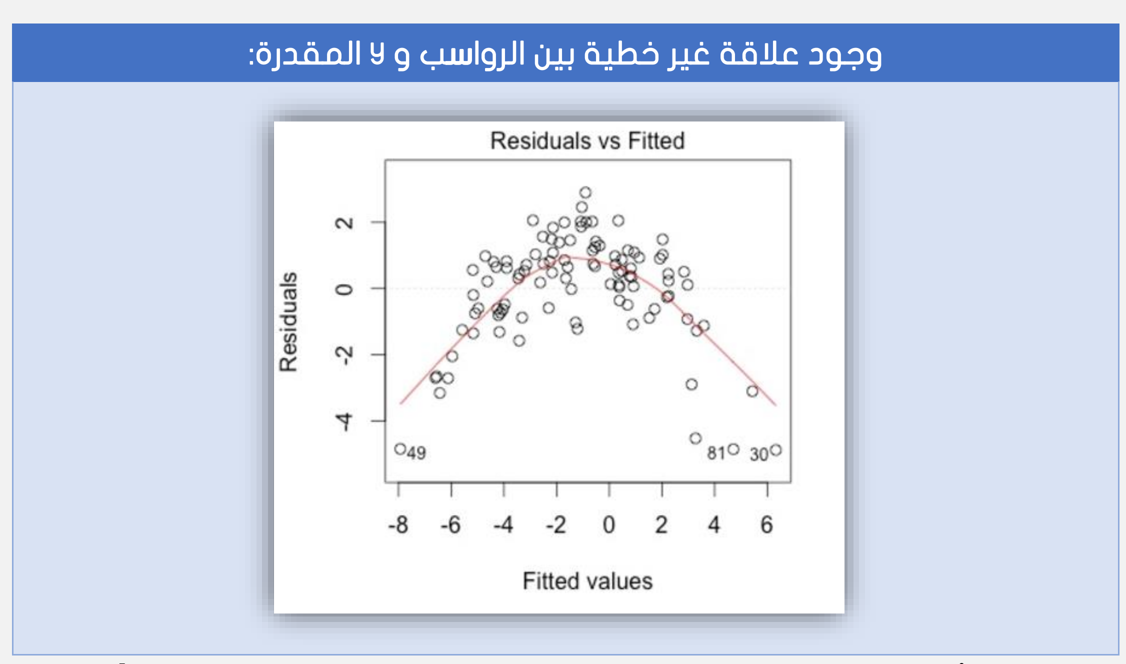


عدم وجود علاقة غير خطية بين الرواسب و y المقدرة:

الرسم يبين عدم وجود أي انتشار للرواسب وفق اتجاه محدد والذي يعني التزام نموذجنا بالفرضية المطلوبة.

أما الشكل الآتي فيبين نموذجاً عن عدم الالتزام بهذه الفرضية:

وجود علاقة غير خطية بين الرواسب و ŷ المقدرة:

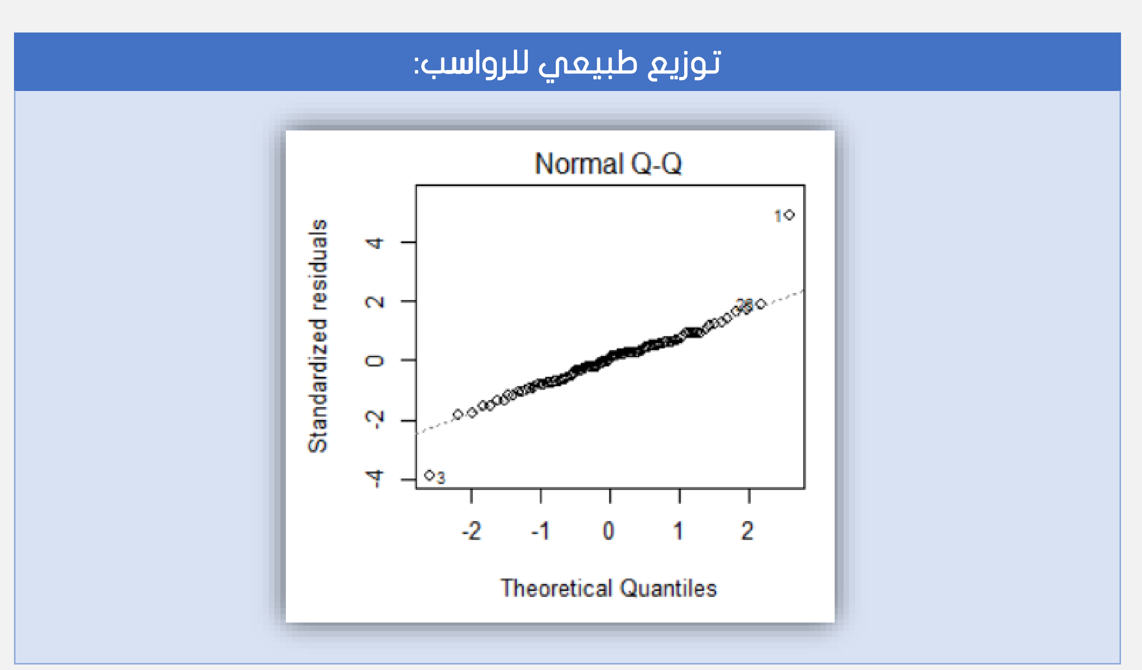


حيث نلاحظ أن العلاقة بين الرواسب وŷ المقدرة لها شكل قطع مكافئ تقريباً.

التوزيع الطبيعي للرواسب (Normality):

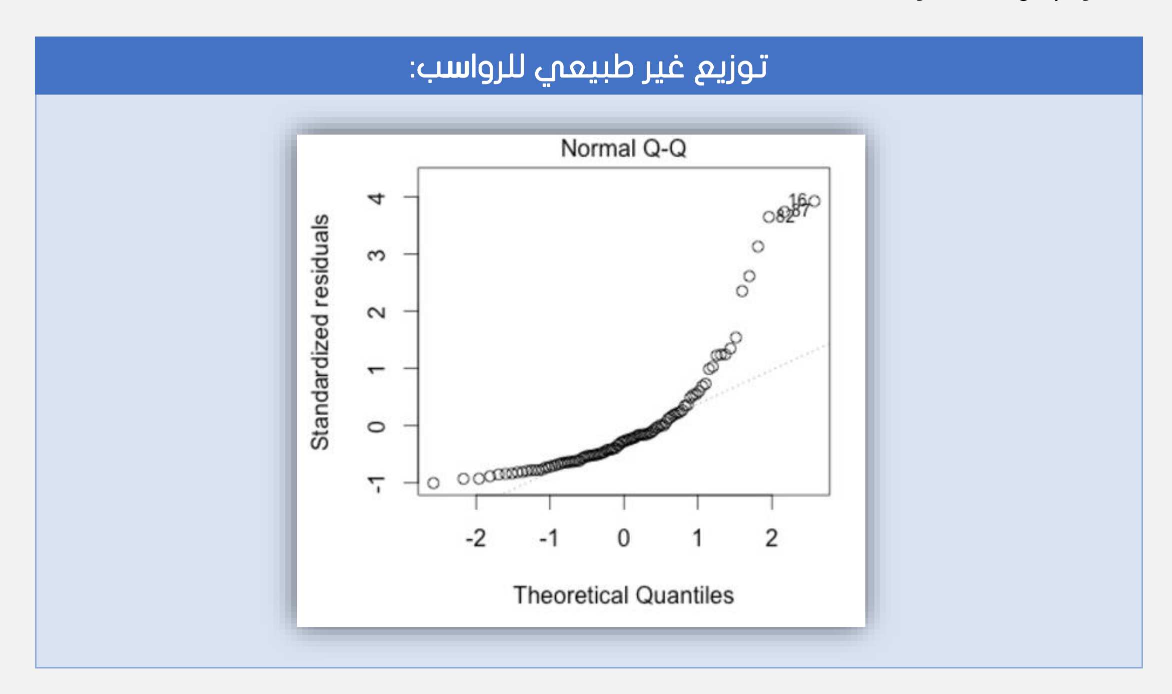
الذي يختبر هذه الفرضية هو الرسم المسمى Normal Q-Q، وحتى تلتزم الرواسب بالتوزيع الطبيعي يجب أن تتجمع النقاط على شكل خط مستقيم دون أن تنحاز عنه بشكل ملحوظ، وفي مثالنا كان لدينا:

توزيع طبيعي للرواسب:



ونلاحظ أن بياناتنا تلتزم بهذه الفرضية بشكل جيد، أما الشكل الآتي فيبين نموذجاً عن عدم الالتزام بهذه الفرضية:

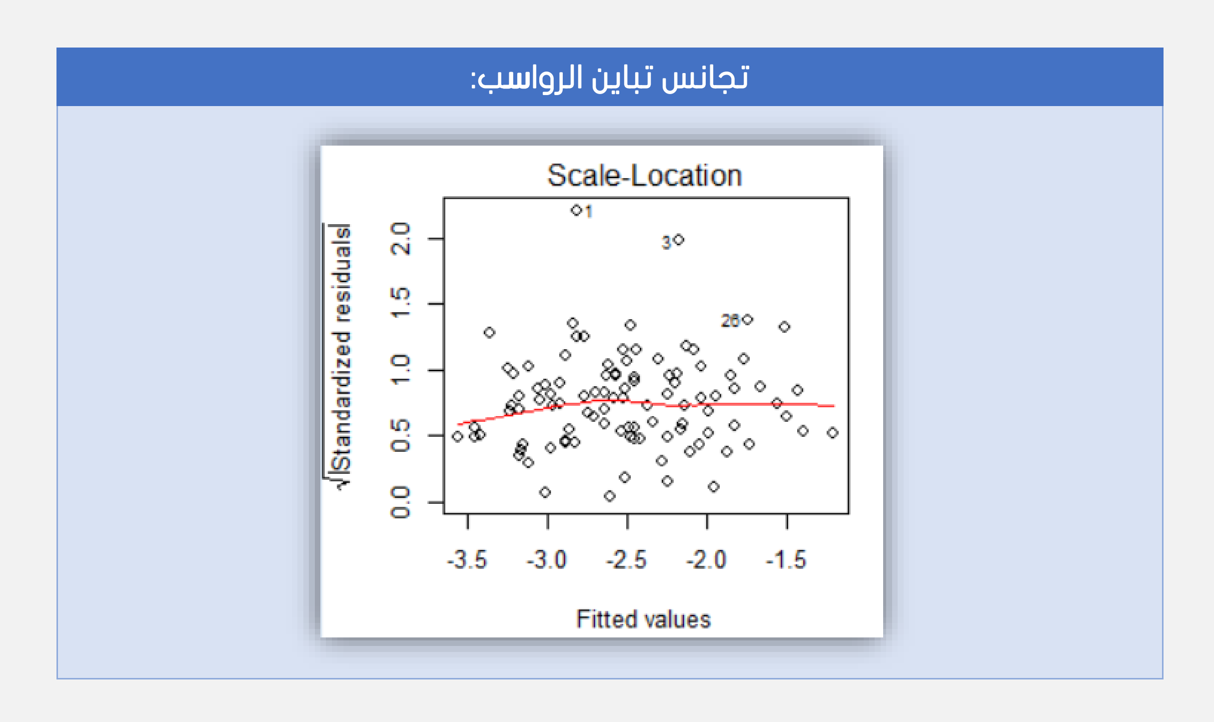
توزيع غير طبيعي للرواسب:



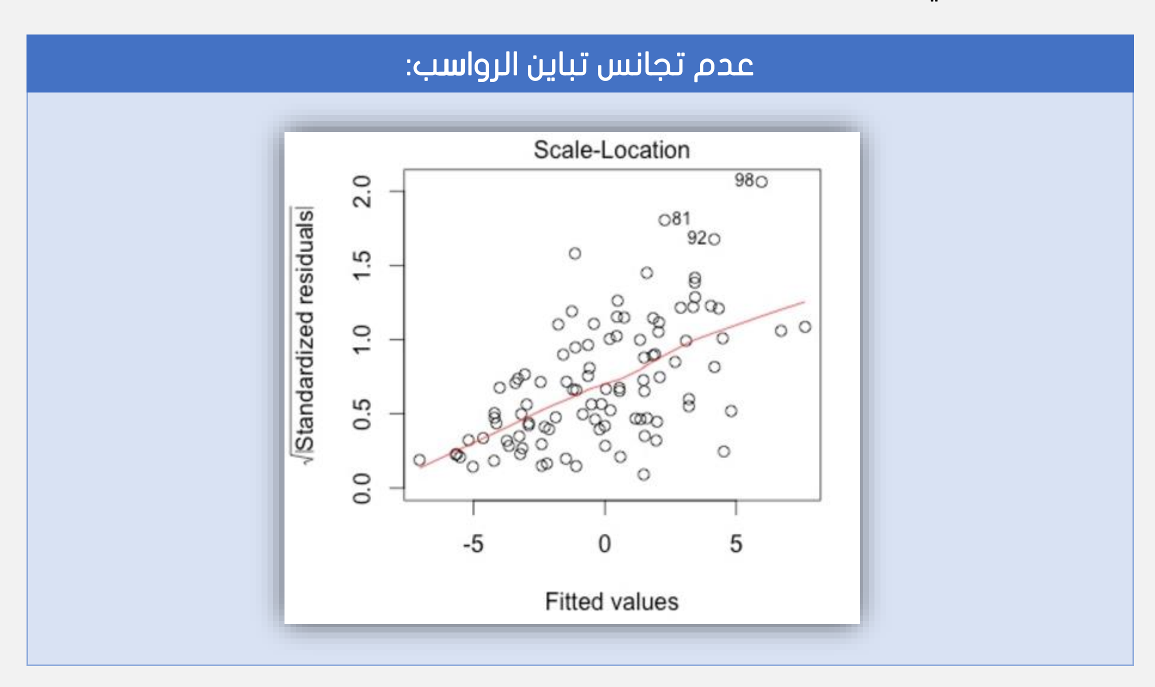
تجانس التباين للرواسب (Homogeneity of Variance):

الذي يختبر هذه الفرضية هو الرسم المسمى Scale Location، فهو يختبر فيما إذا كانت الرواسب تنتشر بشكل متجانس على امتداد قيم المتغيرات المستقلة، فيجب أن يكون الانتشار على شكل مستقيم تتوزع حوله النقاط بتجانس على شكل سيجار تقريباً، أما عند عدم الالتزام بهذه الفرضية فقد نلاحظ أن النقاط تبدأ من الحافة اليسرى السفلى بشكل ضيق ثم تبدأ بالتوسع نحو الزاوية اليمنى العليا، أو تبدأ بشكل متوسع ثم تتضيق.

في مثالنا نجد أن بياناتنا تلتزم بهذه الفرضية حيث نلاحظ من الشكل أدناه أن النقاط تنتشر حول مستقيم على شكل سيجار تقريباً:



أما الشكل الآتي فيبين بيانات لا تلتزم بفرضية تجانس التباين:



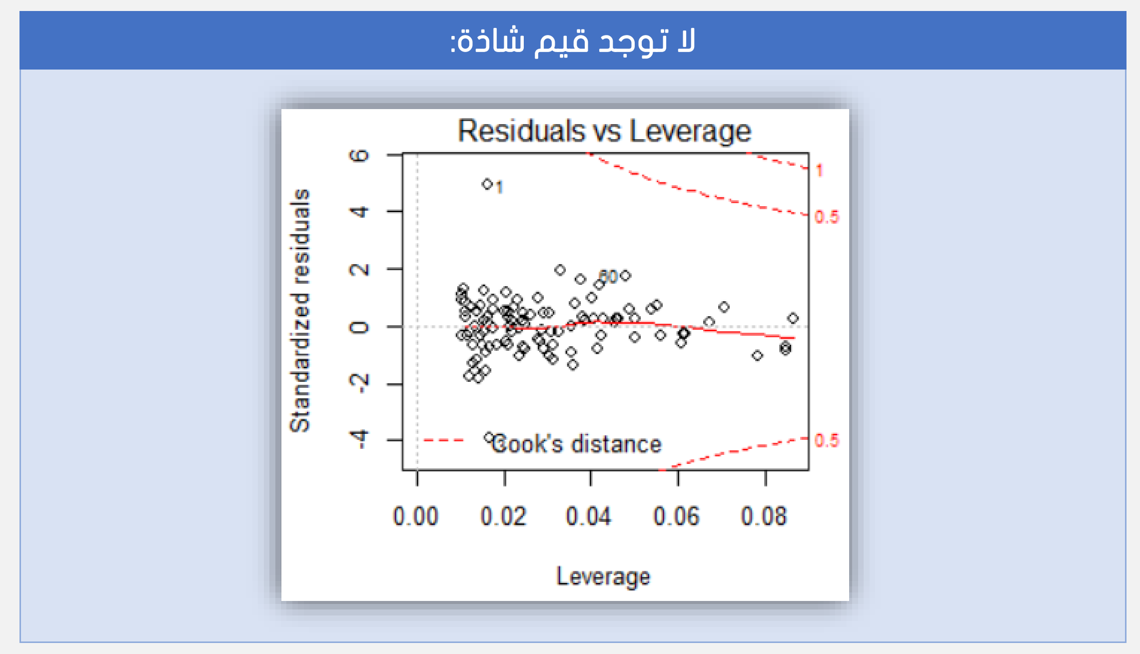
عدم وجود قيم شاذة (Outliers):

الذي يختبر هذا الشرط هو الرسم الأخير Residuals vs Leverage، ففي الانحدار قد نرى من منظورنا أن بعض القيم شاذة، لكنها في الواقع لا تؤثر معنوياً في تقدير خط الانحدار، كما قد نرى أن بعض النقاط ليست شاذة وفي الواقع تكون شاذة بالنسبة لتقدير خط الانحدار كونها تغير الاتجاه العام لهذا الخط، في هذا الرسم سنبحث عن النقاط الشاذة والتي ستتمركز في الزاوية العليا اليمنى أو الزاوية السفلى اليمنى، والتي ستكون خارج الخطوط المنقطة (---) وندعوها بمسافة Cook.

في مثالنا تقع جميع النقاط ضمن مسافة Cook حيث كان لدينا:



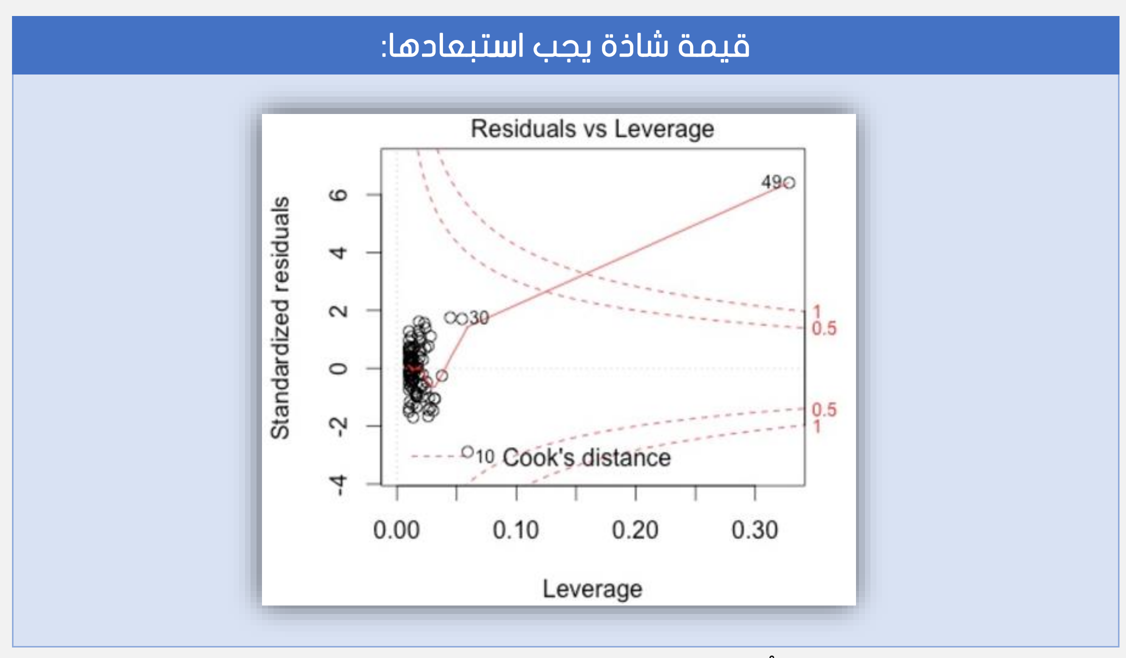
لا توجد قيم شاذة:



قد نشك في القيمة [1] لكن طالما هي لم تتجاوز الخطوط المنقطة فيمكن غض النظر عنها.

أما الشكل الآتي فيبين الحالة التي يكون فيها نقاط شاذة يجب استبعادها:

قيمة شاذة يجب استبعادها:



نلاحظ في الشكل السابق أن النقاط مجمعة بشدة في الزاوية اليسرى وذلك لأن النقطة [49] بعيدة كثيراً عنها، وقد تجاوزت الخطوط المنقطة بكثير، وبالتالي يجب استبعاد معطيات هذه النقطة.

استقلال الرواسب (Independence of Residuals):

يجب ألا تكون الرواسب (الأخطاء) مرتبطة ببعضها البعض في نموذج الانحدار الخطي، والذي يعني أن الخطأ في المستقبل يعتمد على الخطأ في الماضي، ولاختبار هذه الفرضية يستحسن استخدام اختبار Durbin Watson والذي يختبر الفرضية $H_0: \rho = 0$ مقابل $H_1: \rho \neq 0$ وذلك كما يلي:

```
library(car)
durbinWatsonTest(fit)
```

وفي مثالنا لدينا:

التعليمة:

```
durbinWatsonTest(fit)
```

الناتج:

```
lag   Autocorrelation   D-W Statistic   p-value
 1      -0.06482485       1.88327        0.662
Alternative hypothesis: rho != 0
```

وكون p>0.05 فإننا لا نستطيع رفض الفرضية الابتدائية وبالتالي إن شرط استقلال الرواسب محقق.

المصاحبة خطية المتعددة (Multicollinearity):

ويقصد بها ارتباط بعض المتغيرات المستقلة مع بعضها البعض، وهذا غير جائز كوننا ندعوها (متغيرات مستقلة!)، وعادة يستخدم معامل تضخم التباين VIF للتحقق من وجود المصاحبة الخطية المتعددة، فإذا كان VIF>5 لأي من المتغيرات المستقلة يجب علينا حذف هذا المتغير لارتباطه الشديد مع (أحد) أو (باقي) المتغيرات المستقلة الأخرى، يتم حساب VIF باستخدام التعليمة:و

```
vif(fit)
```

وفي مثالنا لدينا:

التعليمة:
vif(fit)
الناتج:
x1 x2 1.010654 1.010654

ونلاحظ أن قيم VIF أصغر من 5 وبالتالي ليست لدينا مشكلة مصاحبة خطية.

حجم العينة (Sample Size):

نظراً للتأثير الكبير لحجم العينة في جودة نموذج الانحدار نشر العالم Green عام 1991 ورقة بحثية حدد فيها حجم العينة المناسب لنموذج الانحدار كالآتي:

1-إذا كان الهدف من النموذج حساب الارتباط المتعدد فقط فإن حجم العينة المناسب هو: $N \geq 50 + 8p$ حيث أن p هو عدد المتغيرات المستقلة.

2-إذا كان الهدف من النموذج معرفة مدى تأثير كل متغير مستقل في المتغير التابع فإن حجم العينة المناسب هو: $N \geq 104 + p$.

الفصل الحادي عشر
الانحدار اللوجستي
(Logistic Regression)

الانحدار اللوجستي هو نموذج يستخدم للتنبؤ باحتمالية وقوع حدث ما وذلك بالاعتماد على متغيرات محددة نسميها المتغيرات المستقلة، والتي يمكن أن تكون كمية أو فئوية. على سبيل المثال: احتمالية حدوث نوبة قلبية عند شخصٍ ما خلال فترة زمنية معينة يمكن التنبؤ بها من خلال معلومات عن عمر المريض وجنسه ومنسب كتلة الجسم لديه. يُستخدم الانحدار اللوجستي بشكل واسع في الطب والعلوم الاجتماعية، كما يستخدم في التسويق لحساب توقعات ميل المستهلك إلى شراء منتج ما أو امتناعه عن الشراء.

ينقسم الانحدار اللوجستي إلى انحدار لوجستي ثنائي (Binary Logistic Regression) وانحدار لوجستي متعدد (Multinomial Logistic Regression).

نكون أمام الانحدار اللوجستي الثنائي عندما يكون للمتغير التابع فئتان مثل التدخين (مدخن / غير مدخن)، الحالة التعليمية (متعلم / غير متعلم)، ...

أما بالانحدار اللوجستي المتعدد فيكون للمتغير التابع أكثر من فئتين، مثل المستوى المعيشي (سيئ / متوسط / جيد)، الوجبة الأهم للفرد (فطور / غداء / عشاء)، ...

سنبدأ أولاً بالانحدار اللوجستي الثنائي:

الانحدار اللوجستي الثنائي (Binary Logistic Regression):

كما سبق وقلنا، إن الانحدار اللوجستي الثنائي يساعدنا على التنبؤ باحتمال الانتماء إلى أحد فئتين لمتغير فئوي Y بالاعتماد على متغيرات $X_1, X_2, \ldots, X_p$ ندعوها المتغيرات المستقلة، ولتطبيق الانحدار اللوجستي الثنائي بعد تخزين بياناتنا في Data frame لها الاسم myData نستخدم التعليمة:

```
blr<-glm(Y~X1+X2+…+Xp,data=myData,family=binomial)
summary(blr)
```

شروط تطبيق الانحدار اللوجستي الثنائي
1. أن يكون المتغير التابع فئوياً وله مستويان فقط. 2. عدم وجود مصاحبة خطية متعددة بين المتغيرات المستقلة. 3. أن يكون هناك علاقة خطية بين المتغيرات المستقلة و log odds حيث نقصد بال Odds Ratio أو ما يسمى بنسب الأرجحية: احتمال تحقق الظاهرة / احتمال عدم تحققها. 4. عدم وجود قيم شاذة. 5. أن يكون حجم العينة كبيراً.

معايير دقة نموذج الانحدار اللوجستي الثنائي (Binary Logistic Regression Model Fit and Accuracy):

في نماذج الانحدار كنا نتفحص قيم $R^2, RMSE$ لتقييم نموذج الانحدار ومدى ملاءمته للبيانات، أما في نماذج الانحدار اللوجستي توجد معايير أخرى سنقوم بسردها:

معيار معلومات أكاكي (Akaike Information Criteria (AIC)):

نستطيع اعتبار معيار AIC بديلاً عن R^2 في نماذج الانحدار، فهو مؤشر هام جداً لملاءمة النموذج ويعتمد القاعدة: كلما كان AIC أصغر دلنا هذا على جودة أكبر للنموذج، ويمتاز معيار AIC بأنه ينقص عند زيادة عدد المتغيرات المستقلة، مما يحل مشكلة الأرقام الوهمية الكبيرة التي تدل على ملاءمة كبيرة (مثل $R^2 = 1$)، لكن مراقبة قيمة AIC وحدها لن تعطينا الفائدة المرجوة، إنما الفائدة تنبع منها للمقارنة بين عدة نماذج لاختيار النموذج الأفضل والذي توافقه قيمة AIC الأصغر.

الانحراف الابتدائي وانحراف الرواسب (Null Deviance and Residual Deviance):

لا تتم الاستفادة من هذين المقياسين إلا بذكرهما معاً، حيث يعتبران مقياسين للخطأ في النموذج، حيث يتم حساب الانحراف الابتدائي أولاً وهو مقياس للخطأ دون إدخال أي متغير مستقل، ثم يتم حساب انحراف الرواسب بإدخال المتغيرات المستقلة وبهذه الحالة يفترض أن يكون الانحراف قد صغر إن كان للمتغيرات المستقلة إسهاماً جيداً في ملاءمة النموذج.

مصفوفة الفوضى (Confusion Matrix):

وهي المعيار الأكثر حسماً وشيوعاً لتقييم نماذج التصنيف ولها الشكل:

	1 (المتنبأة)	0 (المتنبأة)
1 (الفعلية)	الإيجابية الصحيحة TP	السلبية الخاطئة FN
0 (الفعلية)	الإيجابية الخاطئة FP	السلبية الصحيحة TN

ونستطيع استخلاص المقاييس الآتية منها:

الضبط (Accuracy): ويحدد الدقة التنبؤية الكلية للنموذج ويحسب من العلاقة:

Accuracy=(TP+TN)/(TP+FN+FP+TN)

النسبة الإيجابية الصحيحة True Positive Rate (TPR): وتحدد عدد القيم الإيجابية من كل القيم الإيجابية والمصنفة بشكل صحيح وفق العلاقة:

TPR=TP/(TP+FN)

ويعرف أيضاً بالحساسية Sensitivity.

النسبة الإيجابية الخاطئة False Positive Rate (FPR): وتحدد عدد القيم السلبية من كل القيم السلبية والمصنفة بشكل خاطئ وفق العلاقة:

FPR=FP/(FP+TN)

النسبة السلبية الصحيحة True Negative Rate (TNR): وتحدد عدد القيم السلبية من كل القيم السلبية والمصنفة بشكل صحيح وفق العلاقة:

TNR=TN/(TN+FP)

وتعرف أيضاً بالخصوصية Specificity.

النسبة السلبية الخاطئة False Negative Rate (FNR): وتحدد عدد القيم الإيجابية من كل القيم الإيجابية والمصنفة بشكل خاطئ وفق العلاقة:

FNR=FN/(FN+TP)

الدقة Precision: تحدد عدد القيم من جميع القيم التنبؤية الإيجابية والتي هي بالفعل إيجابية وفق العلاقة:

Precision=TP/(TP+FP)

نتيجة F (F Score): وهي المتوسط التوافقي للدقة والحساسية وتتراوح قيمتها بين الصفر والواحد، وكلما اقتربت من الواحد كان النموذج أفضل:

F=2(precision*sensitivity)/(precision+sensitivity)

<u>**مثال:**</u>لتكن لدينا البيانات الآتية والتي تمثل 8 أشخاص مصابين بالسرطان و 8 أشخاص سليمين وعدد السجائر التي يدخنها كل منهم وجنسه:

التشخيص	0	0	0	0	1	0	0	0	1	1	1	1	1	0	1	1
عدد السجائر	10	13	12	12	7	8	12	17	32	37	28	34	32	37	30	28
الجنس	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1

والمطلوب تشكيل نموذج الانحدار اللوجستي لمعرفة تأثير كل من عدد السجائر والجنس على الإصابة بالسرطان.

الحل:

```
cancer<-c(0,0,0,0,1,0,0,0,1,1,1,1,1,0,1,1)
gender<-c(2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1)
noCig<-c(10,13,12,12,7,8,12,17,32,37,28,34,32,37,30,28)
myData<-data.frame("noCig"=noCig,"gender"=gender,"cancer"=cancer)
blr<-glm(cancer~noCig+gender,data=myData,family=binomial)
summary(blr)
```

الناتج:

```
Call:
glm(formula = cancer ~ noCig + gender, family = binomial, data = myData)

Deviance Residuals:
    Min        1Q      Median       3Q       Max
-1.81722  -0.55619   0.04376   0.48011   1.81029

Coefficients:
             Estimate   Std. Error  z value   Pr(>|z|)
(Intercept)  -7.80396    4.31209    -1.810    0.0703 .
noCig         0.18064    0.08391     2.153    0.0313 *
gender        2.55845    1.86681     1.370    0.1705
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 22.181  on 15  degrees of freedom
Residual deviance: 12.772  on 13  degrees of freedom
AIC: 18.772

Number of Fisher Scoring iterations: 5
```

التفسير:

نلاحظ أن المتغير الوحيد الذي يؤدي للإصابة بالسرطان هو عدد السجائر $p=0.0313<0.05$ وإدخال متغير عدد السجائر أبدى تحسيناً في النموذج بسبب أن Residual Deviance أقل من Null Deviance، أما AIC على الرغم من صغرها فلن نستطيع الاستفادة منها لتقييم النموذج لأنه وكما ذكرنا من قبل، تظهر فائدة معيار AIV فقط عند بين المقارنة بين عدة نماذج..

السؤال الآن: وجدنا أن لعدد السجائر دوراً في الإصابة بالسرطان، لكن ما هو هذا الدور وما حجمه؟؟!

نسب الأرجحية (Odds Ratio):

وتعني كم تغير زيادة المتغير المستقل بوحدة واحدة من احتمالية تحقق الظاهرة المدروسة، وهي عبارة عن العدد النيبري مرفوعاً إلى معلمة المتغير في نموذج الانحدار اللوجستي، وفي مثالنا السابق:

مثال:

```
exp(cbind("OR"=coef(blr),confint(blr)))
```

الناتج:

```
                 OR          2.5 %        97.5 %
(Intercept)   4.081169e-04  7.603775e-10  0.205689
noCig         1.197983e+00  1.051239e+00  1.526905
gender        1.291574e+01  6.755416e-01  3116.431011
```

التفسير:

وجدنا سابقاً أن عدد السجائر كان له تأثير معنوي على الإصابة بالسرطان، ومن الجدول السابق نجد أن OR=1.19798 وبالتالي إن زيادة عدد السجائر بسيجارة واحدة يؤدي إلى زيادة احتمالية الإصابة بالسرطان بـ 19.79% وإن 95% فترة ثقة لنسب الأرجحية الموافقة لعدد السجائر هي [1.0512,1.5269].

بإمكاننا أخيراً إيجاد مصفوفة الفوضى بالشكل:

مثال:

```
table(myData$cancer,predict>0.5)
```

الناتج:

```
     FALSE  TRUE
  0    7      1
  1    2      6
```

ومنه نجد دقة التصنيف: 81.25%

الانحدار اللوجستي المتعدد (Multinomial Logistic Regression):

كما سبق وقلنا، إن الانحدار اللوجستي المتعدد يساعدنا على التنبؤ باحتمال الانتماء إلى إحدى فئات متغير فئوي Y (يمتلك أكثر من فئتين) بالاعتماد على متغيرات $X_1, X_2, \dots, X_p$ ندعوها المتغيرات المستقلة، ولتطبيق الانحدار اللوجستي المتعدد بعد تخزين بياناتنا في Data frame لها الاسم myData نستخدم التعليمة:

```
library("nnet")
mlr<-multinom(Y~X1+X2+…+Xp,data=myData)
summary(mlr)
```

شروط تطبيق الانحدار اللوجستي المتعدد
1. أن يكون المتغير التابع فئوياً وله أكثر من مستويين.
2. عدم وجود مصاحبة خطية متعددة بين المتغيرات المستقلة.
3. أن يكون هناك علاقة خطية بين المتغيرات المستقلة و log odds حيث نقصد بال Odds Ratio أو ما يسمى بنسب الأرجحية: احتمال تحقق الظاهرة / احتمال عدم تحققها.
4. عدم وجود قيم شاذة.
5. أن يكون حجم العينة كبيراً.

معايير دقة نموذج الانحدار اللوجستي المتعدد (Multinomial Logistic Regression Model Fit and Accuracy):

وهي نفسها معايير دقة نموذج الانحدار اللوجستي الثنائي.

مثال: لتكن لدينا البيانات الآتية والتي هي عبارة عن 16 شخصاً من مهن مختلفة (1 طبيب 2 مدرس) (3 محامي) ومدة تحملهم للاستفزاز بالساعات ومعدلاتهم الجامعية:

المهنة	1	1	1	1	1	2	2	2	2	2	3	3	3	3	3	3
التحمل	3	2	5	3	3	7	9	6	7	8	8	7	8	11	8	10
المعدل	95	92	97	96	91	80	81	83	86	83	89	75	72	73	71	69

والمطلوب تشكيل نموذج الانحدار اللوجستي للتنبؤ بمهنة الشخص بالاعتماد على معدله ومدى تحمله.

الحل:

```
job<-c(3,3,3,3,3,3,2,2,2,2,2,1,1,1,1,1)
patience<-c(10,8,11,8,7,8,8,7,6,9,7,3,3,5,2,3)
avg<-c(69,71,73,72,75,89,83,86,83,81,80,91,96,97,92,95)
job<-factor(job)
myData<-data.frame("job"=job,"patience"=patience,"avg"=avg)
model<-multinom(job~.,data=myData)
summary(model)
z <- summary(model)$coefficients/summary(model)$standard.errors
p <- (1 - pnorm(abs(z), 0, 1))*2
p
exp(coef(model))
```

الناتج:

```
Call:
multinom(formula = job ~ ., data = myData)

Coefficients:
  (Intercept)   patience      avg
2  -5.178416    8.268146   -0.4500814
3   7.109255    8.842592   -0.6608890

Std. Errors:
  (Intercept)   patience     avg
2   7.656485    29.93532   1.899482
3   7.710670    29.93916   1.902437

Residual Deviance: 9.934945
AIC: 21.93494


  (Intercept)   patience      avg
2  0.4988224   0.7823942   0.8126959
3  0.3565274   0.7677249   0.7282978

    (Intercept)     patience        avg
2  5.636929e-03     3897.715     0.6375763
3  1.223236e+03     6922.914     0.5163920
```

التفسير:

نلاحظ أنه لا يساعد أي من التحمل أو المعدل على التنبؤ بعمل الشخص لأن م المحسوبة في كل الحالات التي تحتها خط أكبر من 0.05، لكن فرضاً لو كانت القيمة الموافقة لـ avg في السطر الثاني هي 0.00 بدلاً من 0.8126959 عندها تفسر هذه القيمة بأن زيادة المعدل بعلامة واحدة تدل على أن احتمال كون الشخص مدرس**اً** أكثر بـ 1/0.6375763=56.84% مرة من كونه طبيباً.

<u>ملاحظة:</u>
نستخدم نسب الأرجحية كما هي إذا كانت أكبر من الواحد، وهي تدل في هذه الحالة على علاقة طردية، بينما نأخذ 1/OR إذا كانت قيمة نسبة الأرجحية أقل من الواحد، وهي تدل في هذه الحالة على علاقة عكسية.

الفصل الثاني عشر
الإحصاء اللامعلمي
(Nonparametric Statistics)

كل ما تم ذكره إلى الآن يندرج تحت ما يسمى الإحصاء المعلمي، ونقصد بالإحصاء المعلمي تلك الأساليب الرياضية التي نستخدمها ونطبقها على بيانات نكون على علم بتوزيعها وتلتزم بشروط محددة، وعند عدم معرفة توزيع البيانات أو عند عدم التزام البيانات بالشروط اللازمة لتطبيق اختبار محدد، ننتقل إلى أساليب أخرى ومنها الإحصاء اللامعلمي.

اختبار ويلكوكسون للعينة الواحدة (One Sample Wilcoxon Test):

وهو البديل اللامعلمي لاختبار t للعينة الواحدة، ويستخدم لاختبار فرق وسيط عينة عن قيمة محددة mu=value ويتم تطبيقه في R بالشكل:

```
wilcox.test(X, mu=value,alt=”two.sided”)
```

مثال:

لتكن لدينا قياسات سكر الدم لمجموعة من المرضى الذين تم علاجهم باستخدام الدواءين Lantus+R، والمطلوب معرفة فيما إذا كان سكر الدم للعينة منضبطاً، علماً أن سكر الدم الطبيعي لمرضى السكر يعتبر 110.

Lantus+R	140	90	110	125	111	128	113	89	110

وقد طبقنا هذا المثال في الفصل التاسع باستخدام اختبار t للعينة الواحدة، لكن الأفضل تطبيق اختبار لامعلمي كون حجم العينة صغيراً جداً، وسنطبق الآن اختبار Wilcoxon للعينة الواحدة:

الحل

```
x<-c(110,89,113,128,111,125,110,90,140)
wilcox.test(x, mu=110,alt="two.sided")
```

الناتج
Wilcoxon signed rank test with continuity correction data: x V = 17, p-value = 0.6726 alternative hypothesis: true location is not equal to 110

نلاحظ من الجدول السابق أن p>0.05 وبالتالي لا يوجد فرق معنوي بين سكر دم المرضى وسكر الدم الطبيعي.

اختبار مان ويتني (Mann Whitney Test):

وهو البديل اللامعلمي لاختبار t للعينتين المستقلتين، ويستخدم لاختبار وجود فروق معنوية في وسيطي مجموعتين مستقلتين، فلو كانت لدينا بيانات مخزنة في متغير تابع اسمه Y ومقسم إلى مجموعتين وفق متغير مستقل اسمه X يمكن اختبار الفرضية: $H_0: median(Y_1) = median(Y_2)$ مقابل $H_1: median(Y_1) \neq median(Y_2)$ بالشكل:

```
wilcox.test(y~x,mu=0,alt="two.sided", paired=F)
```

مثال: لتكن لدينا قياسات سكر الدم لمجموعتين من المرضى حيث تأخذ المجموعة الأولى الدواء Lantus وتأخذ المجموعة الثانية الدواءين Lantus+R:

Lantus	111	128	139	111	121	138	164	149	140
Lantus+R	140	90	110	125	111	128	113	89	110

والمطلوب معرفة أي الدواءين هو الأفضل، هذا المثال نفسه تم حله في الفصل التاسع باستخدام اختبار t والواجب استخدام اختبار Mann Whitney لأن حجم العينة صغير جداً:

الحل

```
y<-c(140,149,164,138,121,111,139,128,111,110,89,113,128,111,125,110,90,140)
x<-c(rep(1,9),rep(2, 9)))
wilcox.test(y~x,mu=0,alt="two.sided", paired=F)
```

الناتج

```
Wilcoxon rank sum test with continuity correction

data:  y by x
W = 65, p-value = 0.03342
alternative hypothesis: true location shift is not equal to 0
```

التفسير:

بما أن p<0.05 بالتالي يوجد فروق معنوية بين العلاجين عند مستوى المعنوية 5%.

اختبار ويلكوكسون لعينتين مرتبطتين (Paired Samples Wilcoxon Test):

وهو البديل اللامعلمي لاختبار t للعينتين المرتبطتين (العينة المزدوجة)، ويستخدم لاختبار وجود فروق معنوية في وسيطي مجموعة في ظرفين مختلفين، فلو رمزنا للعينة في الظرف الأول = بـ X وللعينة في الظرف الثاني بـ Y عندها يمكن اختبار الفرضية: $H_0: median(X) = median(Y)$ مقابل $H_1: median(X) \neq median(Y)$ بالشكل:

```
wilcox.test(X,Y,mu=0,alt="two.sided", paired=T)
```

<u>مثال:</u> قمنا بتطبيق نظام حمية على عينة من النساء لمدة شهر وقمنا بتسجيل أوزانهن قبل وبعد الحمية فكانت النتائج كما يلي:

قبل	65	66	62	59	62	74	63	69	65
بعد	60	60	59	58	54	67	58	62	60

والمطلوب معرفة فيما إذا كانت الحمية مجدية، هذا المثال نفسه تم حله في الفصل التاسع باستخدام اختبار t والواجب استخدام اختبار Wilcoxon لأن حجم العينة صغير جداً:

الحل

```
x<-c(65,69,63,74,62,59,62,66,65)
y<-c(60,62,58,67,54,58,59,60,60)
wilcox.test(x,y,mu=0,alt="two.sided", paired=T)
```

الناتج

```
        Wilcoxon signed rank test with continuity correction

data:  x and y
V = 45, p-value = 0.008849
alternative hypothesis: true location shift is not equal to 0
```

التفسير:

بما أن p<0.05 بالتالي يوجد فروق معنوية بأوزان النساء قبل الحمية وبعد الحمية عند مستوى المعنوية 5%.

اختبار كروسكال واليز (Kruskal Wallis Test):

وهو البديل اللامعلمي لاختبار One Way ANOVA، أي أنه يستخدم لدراسة وجود فروق معنوية بين عدة مجموعات، ويمكن تطبيقه بالشكل:

```
kruskal.test(y~x,alt="two.sided")
```

مثال:

لدراسة الاختلاف في معدلات الطلاب باختلاف طريقة التدريس كانت لدينا النتائج الآتية:

A	B	C
70	89	88
73	90	82
72	95	87
74	77	89
75	79	83
73	82	85
71	79	89

حيث تمثل A طريقة التدريس التقليدية، و B التدريس مع جهاز إسقاط، و C التدريس مع مذاكرات دورية.

الحل

```
y<-c(70,73,72,74,75,73,71,89,90,95,77,79,82,79,88,82,87,89,83,85,89)
x<-rep(c(1,2,3),each=7)
kruskal.test(y~x)
```

الناتج

```
        Kruskal-Wallis rank sum test

data:  y by x
Kruskal-Wallis chi-squared = 13.544, df = 2, p-value = 0.001145
```

والذي يهمنا من الناتج هو القيمة p=0.001145 والتي هي أقل من 0.05 وبالتالي نستنتج أن طريقتين على الأقل من الطرائق الثلاثة تختلفان عن بعضهما البعض اختلافاً معنوياً، وبالتالي نسأل السؤال الآتي: أي الطرائق هي التي تختلف عن بعضها البعض؟

اختبار ويلكوكسون للمقارنات المتعددة وتصحيح هولم لقيمة المعنوية (Pairwise Wilcoxon and Holm Adjusted P-Value):

يستخدم الاختبار السابق بعد تطبيق اختبار Kruskal Wallis وملاحظة أن $p<0.05$ وذلك لمعرفة أي المجموعات هي التي تختلف عن بعضها البعض اختلافاً معنوياً، وذلك بالشكل:

```
pairwise.wilcox.test(y,x,paired=F,p.adj="holm")
```

وفي مثالنا نجد:

الحل

```
pairwise.wilcox.test(y,x,paired=F,p.adj="holm")
```

الناتج

```
        Pairwise comparisons using Wilcoxon rank sum test

data:  y and x

      1        2
2   0.0063     -
3   0.0063   0.6526

P value adjustment method: holm
```

وبالتالي إن الطريقة الأولى تختلف عن كل من الطريقة الثانية والثالثة معنوياً عند مستوى المعنوية 5%.

توجد العديد من البدائل اللامعلمية التي يمكن الاطلاع عليها في الكتب المختصة وسنكتفي في كتابنا بهذا القدر.

الفصل الثالث عشر
السلاسل الزمنية
(Time Series)

سنستعرض في هذا الفصل طريقة إجراء أهم التحليلات الأساسية للسلاسل الزمنية، وأول ما عليك فعله هو قراءة بيانات السلسلة الزمنية:

قراءة بيانات السلسلة الزمنية (Time Series Data Reading):

سنعمل في هذا الفصل على بيانات جاهزة وسنبدأ بأعمار وفاة 42 ملكاً ناجحاً لبريطانيا سنقوم باستيرادها من الرابط https://robjhyndman.com/tsdldata/misc/kings.dat، وسنتجاهل أول ثلاثة أسطر كونها تحتوي عبارات توضيحية حول البيانات:

```
kings <- scan("https://robjhyndman.com/tsdldata/misc/kings.dat",skip=3)
```

والخطوة الآتية هي تحويل البيانات السابقة إلى سلسلة زمنية (أي ربطها بالزمن):

التعليمة
tskings <- ts(kings) tskings
الناتج
Time Series: Start = 1 End = 42 Frequency = 1 [1] 60 43 67 50 56 42 50 65 68 43 65 34 47 34 49 41 13 35 53 56 16 43 69 59 48 59 86 55 68 51 33 49 67 77 81 67 71 81 68 [40] 70 77 56

قمنا باستخدام التابع ()ts لتحويل البيانات إلى بيانات سلسلة زمنية سنوية، إما إذا أردنا تحويلها إلى سلسلة زمنية شهرية بإمكاننا وضع الوسيط frequency=12 في التابع ()ts كما بإمكاننا جعلها سلسلة زمنية ربع سنوية بوضع الوسيط frequency=4، بإمكاننا أيضاً تحديد تاريخ بداية السلسلة الزمنية باستخدام الوسيط start=c(year,month/quarter) كما في المثال الآتي لبيانات أخرى تمثل عدد الولادات الشهرية في New York من 1946 حتى 1959 من الموقع: https://robjhyndman.com/tsdldata/data/nybirths.dat .

التعليمة

```
births <- scan("https://robjhyndman.com/tsdldata/data/nybirths.dat")
tsbirths <- ts(births, frequency=12, start=c(1946,1))
tsbirths
```

الناتج

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1946	26.663	23.598	26.931	24.740	25.806	24.364	24.477	23.901	23.175	23.227	21.672	21.870
1947	21.439	21.089	23.709	21.669	21.752	20.761	23.479	23.824	23.105	23.110	21.759	22.073
1948	21.937	20.035	23.590	21.672	22.222	22.123	23.950	23.504	22.238	23.142	21.059	21.573
1949	21.548	20.000	22.424	20.615	21.761	22.874	24.104	23.748	23.262	22.907	21.519	22.025
1950	22.604	20.894	24.677	23.673	25.320	23.583	24.671	24.454	24.122	24.252	22.084	22.991
1951	23.287	23.049	25.076	24.037	24.430	24.667	26.451	25.618	25.014	25.110	22.964	23.981
1952	23.798	22.270	24.775	22.646	23.988	24.737	26.276	25.816	25.210	25.199	23.162	24.707
1953	24.364	22.644	25.565	24.062	25.431	24.635	27.009	26.606	26.268	26.462	25.246	25.180
1954	24.657	23.304	26.982	26.199	27.210	26.122	26.706	26.878	26.152	26.379	24.712	25.688
1955	24.990	24.239	26.721	23.475	24.767	26.219	28.361	28.599	27.914	27.784	25.693	26.881
1956	26.217	24.218	27.914	26.975	28.527	27.139	28.982	28.169	28.056	29.136	26.291	26.987
1957	26.589	24.848	27.543	26.896	28.878	27.390	28.065	28.141	29.048	28.484	26.634	27.735
1958	27.132	24.924	28.963	26.589	27.931	28.009	29.229	28.759	28.405	27.945	25.912	26.619
1959	26.076	25.286	27.660	25.951	26.398	25.565	28.865	30.000	29.261	29.012	26.992	27.897

رسم السلاسل الزمنية (Plotting Time Series):

يمكن رسم السلسلة الزمنية timeSeries باستخدام التعليمة:

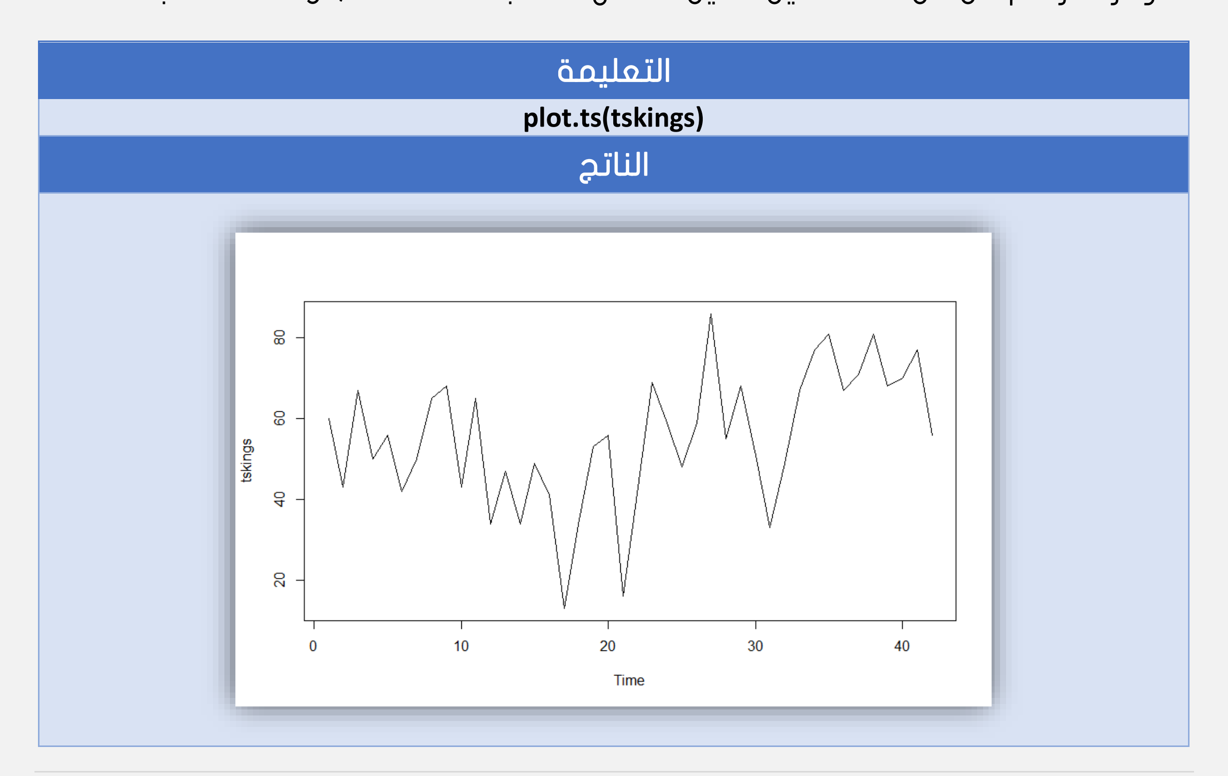
```
plot.ts(timeSeries)
```

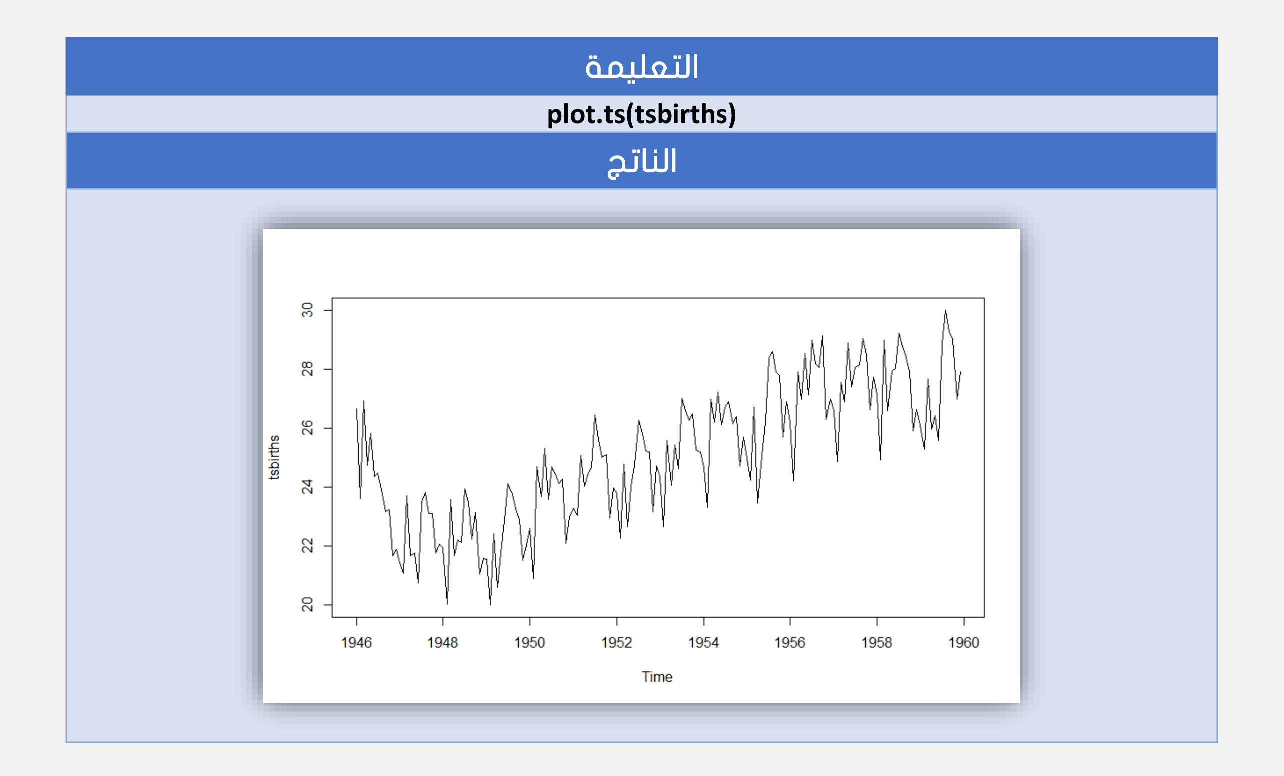
فلو أردنا رسم كل من السلسلتين اللتين أنشأنهما سابقاً tskings,tsbirths نكتب:

التعليمة

```
plot.ts(tskings)
```

الناتج





تفكيك السلاسل الزمنية (Decomposing Time Series):

ونقصد بتفكيك السلسلة الزمنية عزل المركبات المكونة لها وهي مركبة الاتجاه العام والمركبة الموسمية ومركبة الخطأ العشوائي، وسنفترض أن سلسلتنا مشكلة من هذه المركبات بواسطة الجمع، أي أن: $TS = Trend + Seasonal + Random$، وإذا كانت السلسلة مشكلة من هذه المركبات بواسطة الجداء يمكن ردها إلى حالة الجمع بأخذ اللوغاريتم لبيانات السلسلة الزمنية.

لتفكيك سلسلة زمنية يجب علينا أولاً تنزيل الحزمة TTR بالشكل:

```
install.packages("TTR")
library("TTR")
```

بفرض أن سلسلتنا **لا تمتلك مركبة موسمية** عندها يمكن التخلص من المركبة العشوائية والحصول على مركبة الاتجاه العام باستخدام المتوسطات المتحركة البسيطة من المرتبة span بواسطة التعليمة:

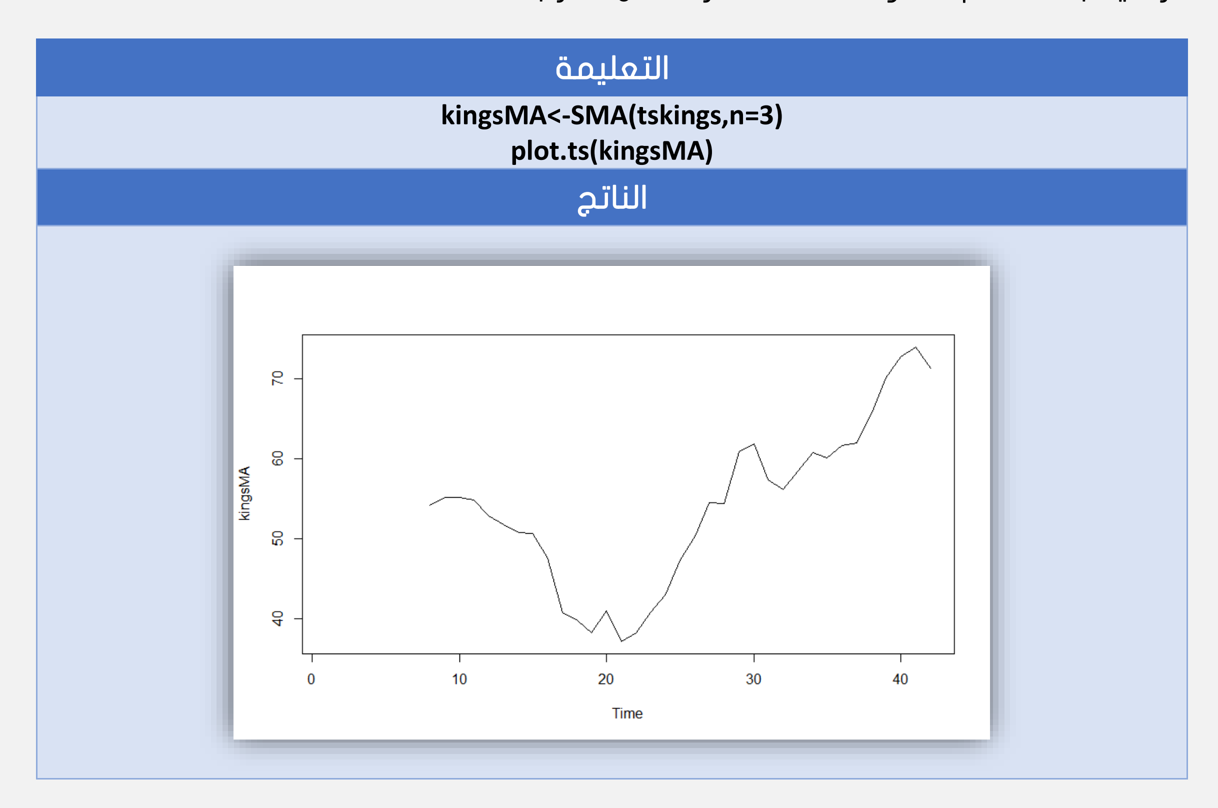
```
SMA(timeSeries,n=span)
```

بالعودة لسلسلة أعمار وفاة ملوك بريطانيا والتي يبدو من رسمها البياني أنها لا تمتلك مركبة موسمية وبالتالي يمكن الحصول على مركبة الاتجاه العام بتنعيم السلسلة الزمنية باستخدام المتوسطات المتحركة من المرتبة الثامنة مثلاً:

التعليمة

```
kingsMA<-SMA(tskings,n=3)
plot.ts(kingsMA)
```

الناتج



والرسم السابق يسمح لنا بالقول: إن أعمار وفاة الملوك انخفضت من 55 سنة إلى 38 سنة تقريباً خلال أول 20 سنة ثم بدأت بالزيادة فيما بعد إلى 73 سنة للملك رقم 40 تقريباً.

أما إذا كانت سلسلتنا موسمية عندها نستخدم التعليمة:

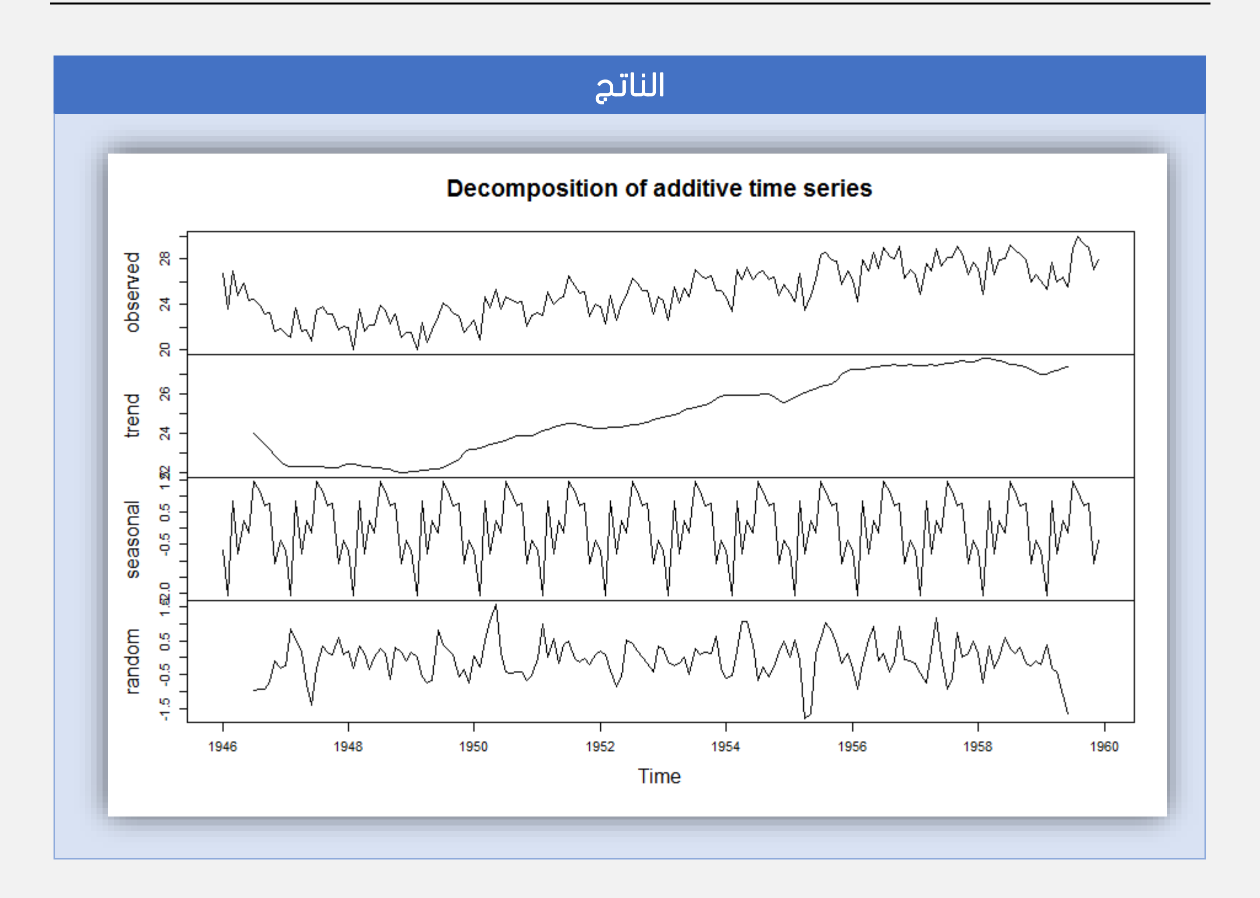
```
decompose(mySeries)
```

فبالنسبة لبيانات سلسلة الولادات في New York:

التعليمة

```
birthsComponents<-decompose(tsbirths)
plot(birthsComponents)
```

الناتج



يمكننا الوصول لأي مركبة من مركبات السلسلة الزمنية بذكر اسم السلسلة ثم الرمز $ ثم اسم المركبة كما يلي:

مركبة الاتجاه العام	timeSeries$trend
المركبة الموسمية	timeSeries$seasonal
مركبة الخطأ العشوائي	timeSeries$random

التمهيد الأسي البسيط (Simple Exponential Smoothing):

إذا كانت لديك سلسلة زمنية لا تمتلك مركبة موسمية وليس لها اتجاه عام عندها بإمكانك إجراء تنبؤات باستخدام التمهيد الأسي البسيط، والذي يقوم بالتنبؤ بالقيم اللاحقة بالاعتماد على القيم السابقة بوزنها بوسيط alpha يأخذ قيمه بين الصفر والواحد، فكلما كانت قيمة alpha أقرب إلى الواحد دلنا هذا على ارتباط أكبر للقيمة اللاحقة بالقيمة السابقة، ويمكن إجراء التمهيد الأسي البسيط باستخدام التعليمة:

```
HoltWinters(timeSeries,beta=F,gamma=F)
```

سنطبق التمهيد الأسي البسيط على بيانات كميات هطول الأمطار السنوية في لندن خلال العام 1813-1912 من الموقع https://robjhyndman.com/tsdldata/hurst/precip1.dat:

التعليمة

```
rain <- scan("https://robjhyndman.com/tsdldata/hurst/precip1.dat",skip=1)
raints <- ts(rain,start=c(1813))
rainforecasts <- HoltWinters(raints, beta=FALSE, gamma=FALSE)
rainforecasts
plot(rainforecasts)
```

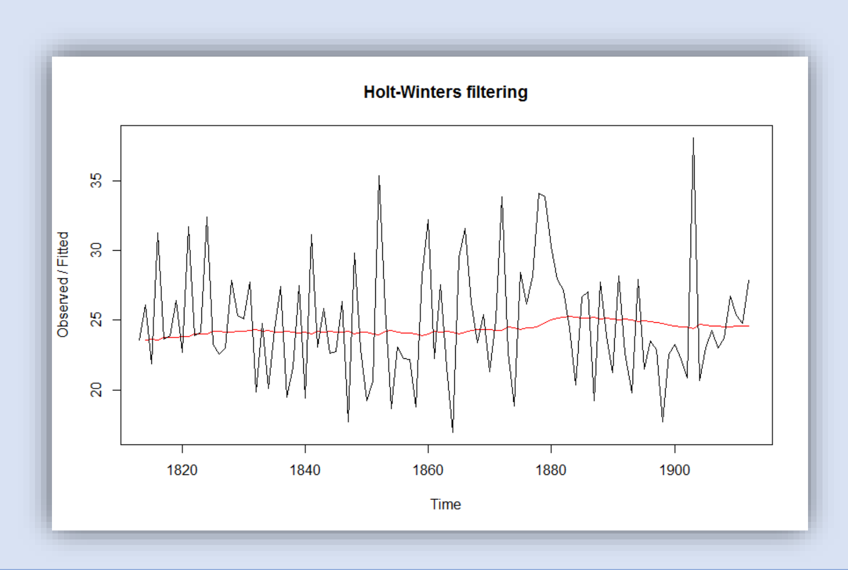
الناتج

```
Holt-Winters exponential smoothing without trend and without seasonal component.

Call:
HoltWinters(x = raints, beta = FALSE, gamma = FALSE)

Smoothing parameters:
 alpha: 0.02412151
 beta : FALSE
 gamma: FALSE

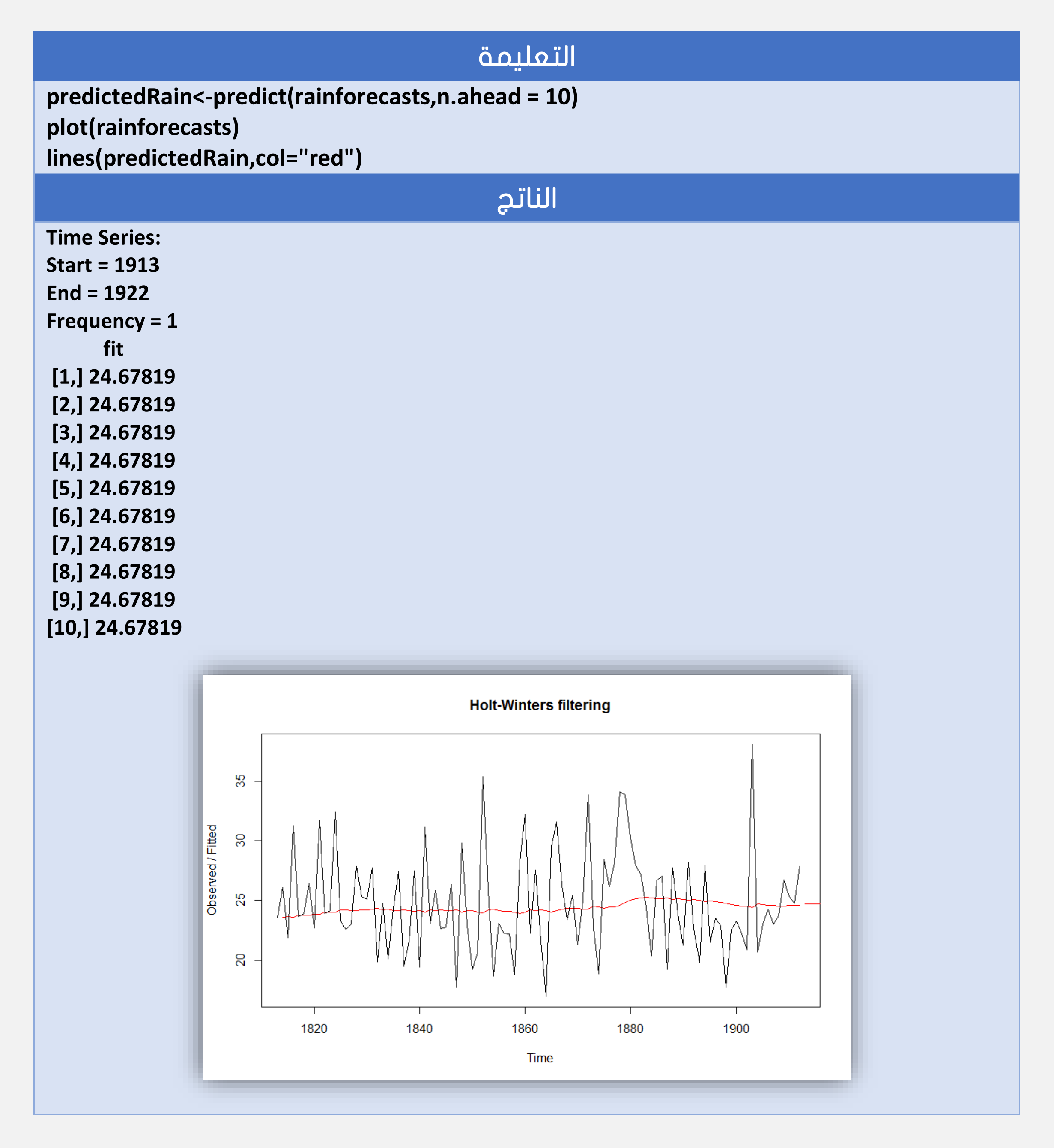
Coefficients:
     [,1]
a 24.67819
```



وللتنبؤ بـ m سنة لاحقة بالاعتماد على سلسلة التنبؤات tsForecasts المبنية بواسطة التابع HoltWinters نستخدم التعليمة:

```
predict(tsForecasts,n.ahead=m)
```

بالعودة لمثالنا السابق وللتنبؤ بسلسلة الأمطار لعشر سنوات لاحقة نكتب:



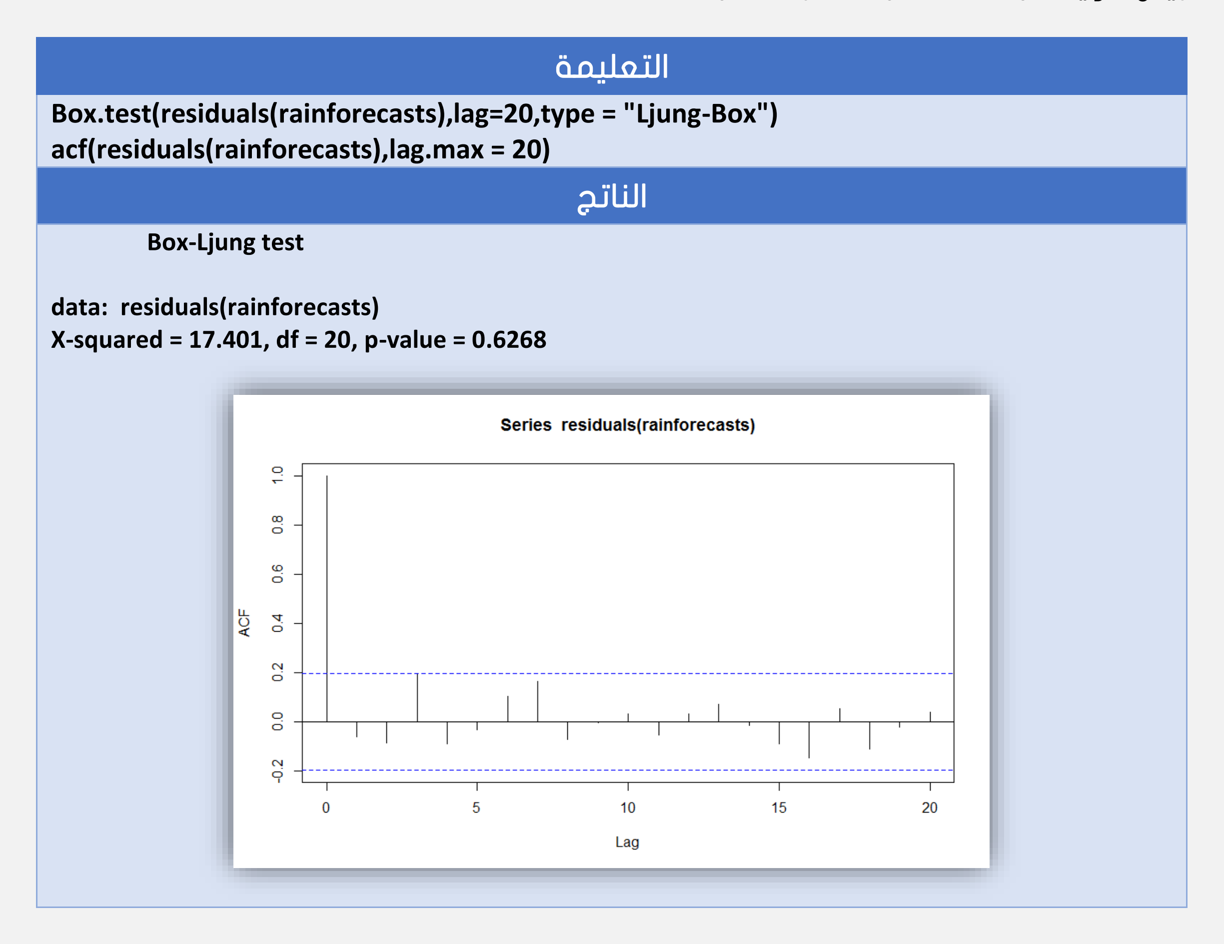
التعليمة
predictedRain<-predict(rainforecasts,n.ahead = 10) plot(rainforecasts) lines(predictedRain,col="red")
الناتج
Time Series: Start = 1913 End = 1922 Frequency = 1 fit [1,] 24.67819 [2,] 24.67819 [3,] 24.67819 [4,] 24.67819 [5,] 24.67819 [6,] 24.67819 [7,] 24.67819 [8,] 24.67819 [9,] 24.67819 [10,] 24.67819

حتى نحكم على جودة نموذجنا يمكننا حساب مجموع مربعات الأخطاء SSE بالشكل:

التعليمة
rainforecasts$SSE
الناتج
[1] 1828.855

ولن نستطيع الاستفادة من هذه القيمة وحدها، إنما يمكن استخدامها للمقارنة بين عدة نماذج لاختيار النموذج الأفضل والذي يكون فيه SSE أصغر ما يمكن.

وهناك شرط هام يجب التأكد منه لنعلم فيما إذا كان نموذج التمهيد الأسي ملائماً لبياناتنا أو يجب البحث عن نموذج آخر، وهو استقلال الرواسب وعدم وجود ارتباط ذاتي بينها، ويمكن التأكد من ذلك بالشكل:



التعليمة

```
Box.test(residuals(rainforecasts),lag=20,type = "Ljung-Box")
acf(residuals(rainforecasts),lag.max = 20)
```

الناتج

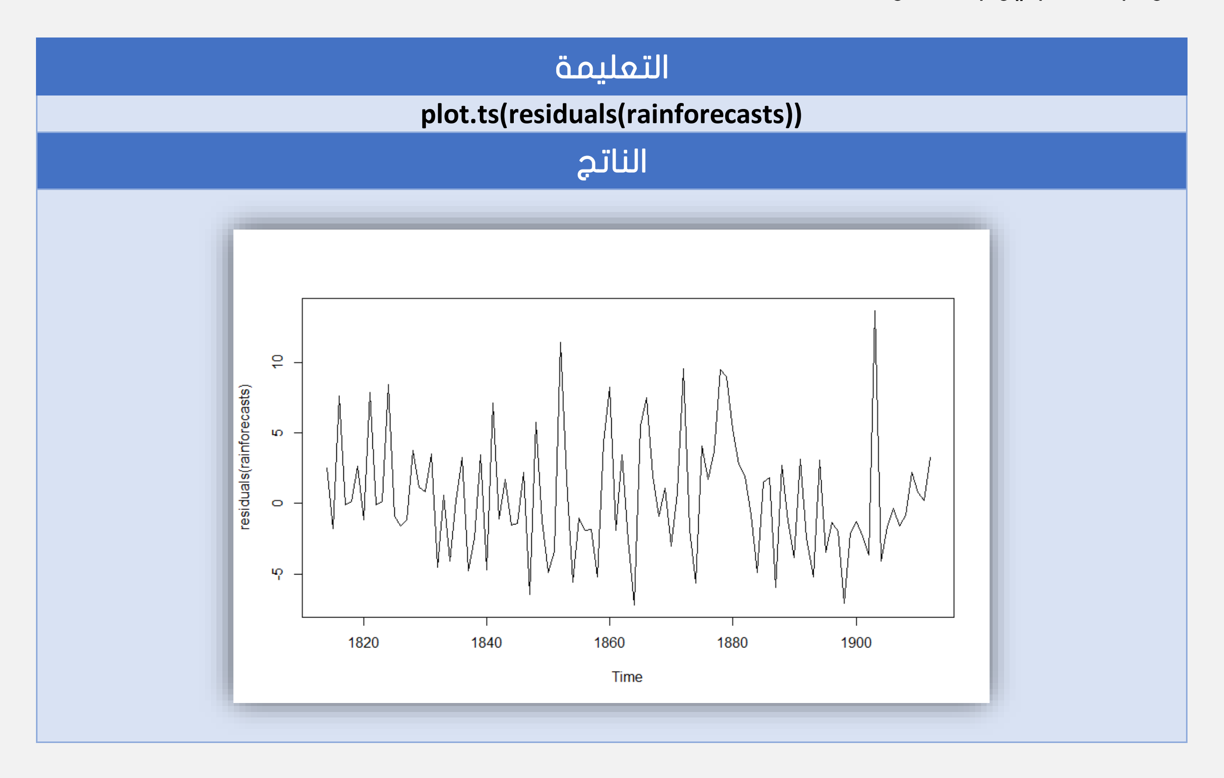
```
        Box-Ljung test

data:  residuals(rainforecasts)
X-squared = 17.401, df = 20, p-value = 0.6268
```


يستخدم اختبار Ljung-Box لاختبار صحة الفرضية الابتدائية القائلة بعدم وجود ارتباط ذاتي بين الرواسب، فإذا كان p>0.05 قبلنا هذه الفرضية أما إذا كان p<0.05 رفضناها، وفي مثالنا p=0.6268>0.05 وبالتالي لا يوجد ارتباط ذاتي بين الرواسب.

كما يقدم لنا الرسم البياني أيضاً وسيلة أخرى لاختبار الفرضية السابقة، حيث يجب أن لا تتجاوز قيم الارتباط الذاتي خطوط المعنوية المنقطة، وفي مثالنا تقترب قيمة الارتباط الذاتي عند الفجوة 3 فقط من خط المعنوية دون أن تتجاوزه، وبالتالي إن الرسم السابق يدعم النتائج التي توصلنا إليها باستخدام اختبار Ljung-Box.

يجب أيضاً التأكد من أن للرواسب توزيعاً طبيعياً بمتوسط صفري وتباين ثابت، ويمكن التأكد من ثبات التباين بالشكل:



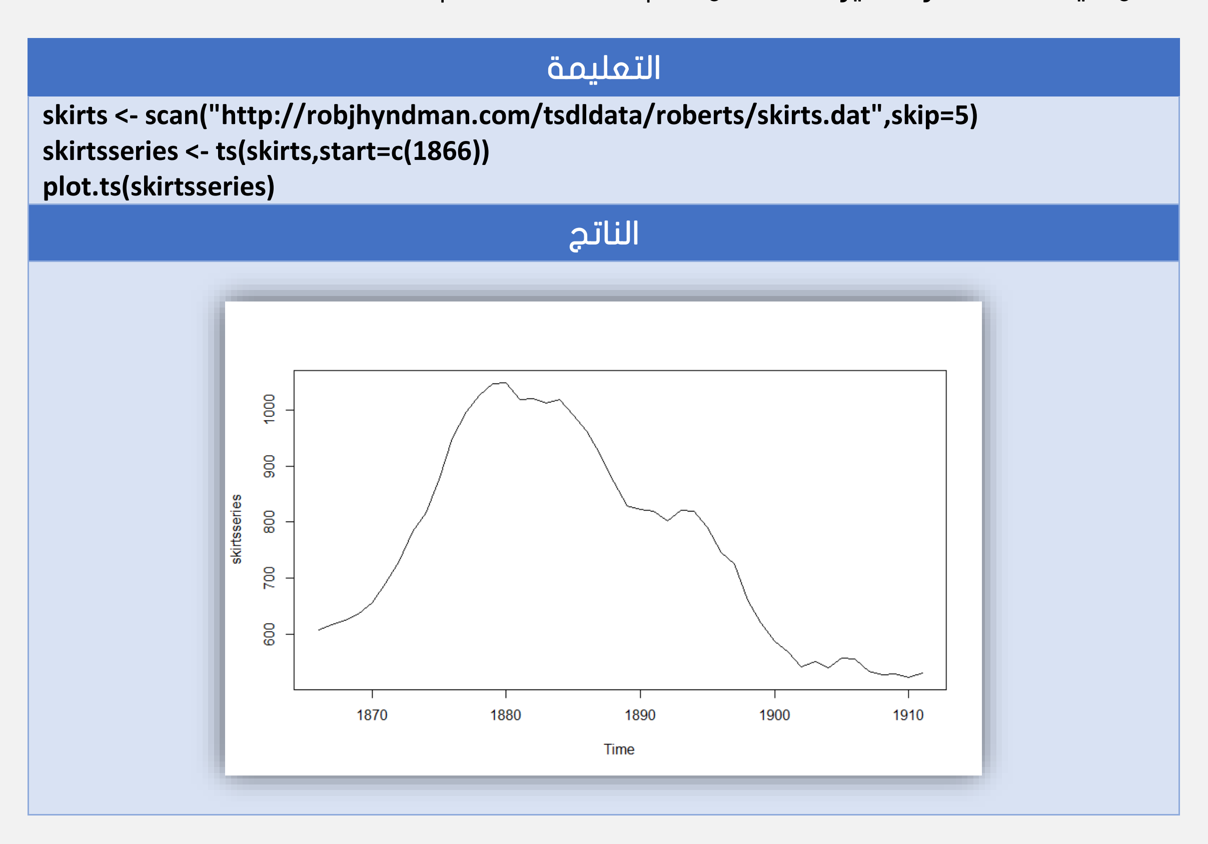
ونلاحظ أن تباين الرواسب ثابت عبر الزمن لأن الرواسب تنتشر بشكل عشوائي حول خط وهمي ليس له اتجاه عام في الزيادة أو النقصان.

للتأكد من أن للرواسب توزيعاً طبيعياً بمتوسط صفري نستطيع تطبيق أي اختبار من اختبارات الطبيعية التي ذكرناها في الفصل السابع، وكذلك نستطيع التأكد من أن للرواسب متوسطاً صفرياً باستخدام اختبار t للعينة الواحدة، وسنترك هذين الموضوعين للقارئ.

التمهيد الأسي لهولت (Holt's Exponential Smoothing):

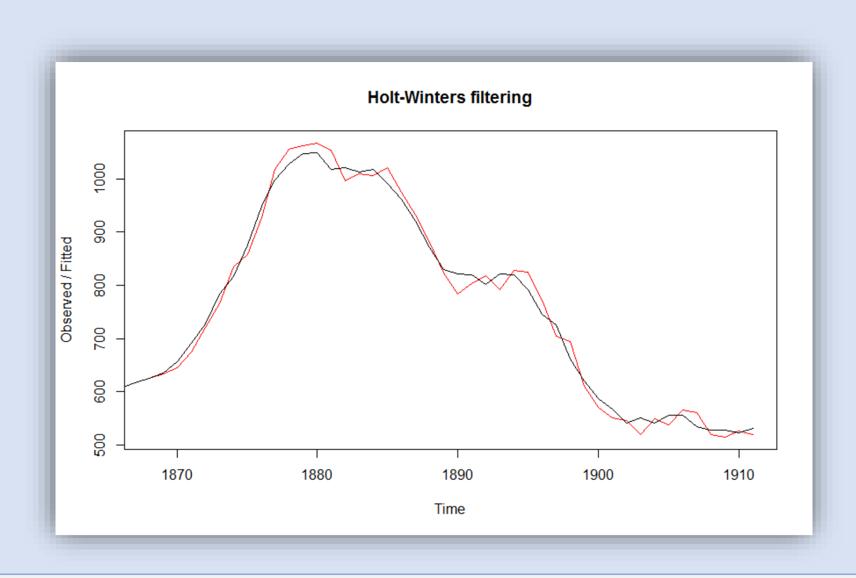
يمكن استخدام نموذج هولت للتمهيد الأسي إذا كان للسلسلة الزمنية اتجاهاً عاماً بالزيادة أو النقصان دون مركبة موسمية، سيكون لدينا في نموذج هولت معلمتين alpha و beta حيث تمثل alpha معلمة المستوى وتمثل beta معلمة الاتجاه العام، ولكل من المعلمتين السابقتين قيم بين 0 و 1.

سنستخدم البيانات الآتية: https://robjhyndman.com/tsdldata/roberts/skirts.dat والتي تمثل قياسات الخصر لتنانير النساء من عام 1866 حتى عام 1911:

التعليمة
skirts <- scan("http://robjhyndman.com/tsdldata/roberts/skirts.dat",skip=5) skirtsseries <- ts(skirts,start=c(1866)) plot.ts(skirtsseries)
الناتج


نلاحظ أن بيانات السلسلة الزمنية لها اتجاه عام بالزيادة من عام 1866 حتى 1880 ثم لها اتجاه عام بالتناقص حتى 1911.

سنستخدم التابع ()HoltWinters لنمذجة السلسلة الزمنية السابقة بالشكل:

التعليمة
skirtsforecasts<-HoltWinters(skirtsseries,gamma=FALSE) skirtsforecasts plot(skirtsforecasts)
الناتج
Holt-Winters exponential smoothing with trend and without seasonal component. Call: HoltWinters(x = skirtsseries, gamma = FALSE) Smoothing parameters: alpha: 0.8383481 beta : 1 gamma: FALSE Coefficients: [,1] a 529.308585 b   5.690464 

ويمكن التنبؤ بالقيم المستقبلية للسلسلة الزمنية بنفس الطريقة المتبعة في نموذج التمهيد الأسي البسيط.

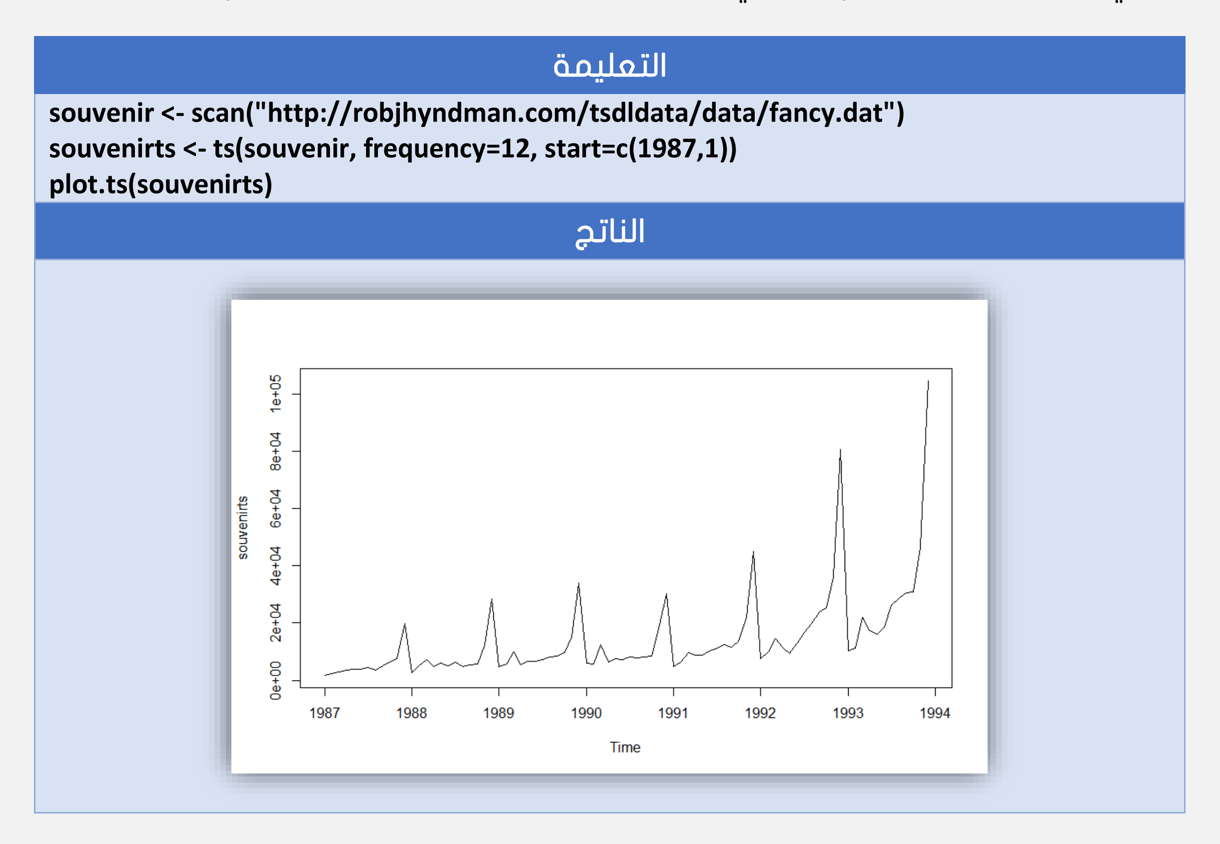
كذلك يجب التأكد من عدم وجود ارتباط ذاتي للرواسب، ومن أن الرواسب تتوزع وفق التوزيع الطبيعي بمتوسط صفري، وكل هذا سنتركه للقارئ لأنه يتم تماماً كما قد سبق في التمهيد الأسي البسيط.

التمهيد الأسي لهولت ووينترز (Holt-Winters Exponential Smoothing):

نستخدم التمهيد الأسي لهولت ووينترز إذا كان للسلسلة الزمنية اتجاهاً عاماً ومركبة موسمية، وبالتالي سيكون لنموذجنا ثلاث معلمات، alpha معلمة المستوي، و beta معلمة الاتجاه العام، و gamma معلمة الموسمية.

سنستخدم بيانات السلسلة الزمنية: http://robjhyndman.com/tsdldata/data/fancy.dat
والتي تمثل المبيعات الشهرية في أحد شواطئ أستراليا من 1987 حتى نهاية 1993:

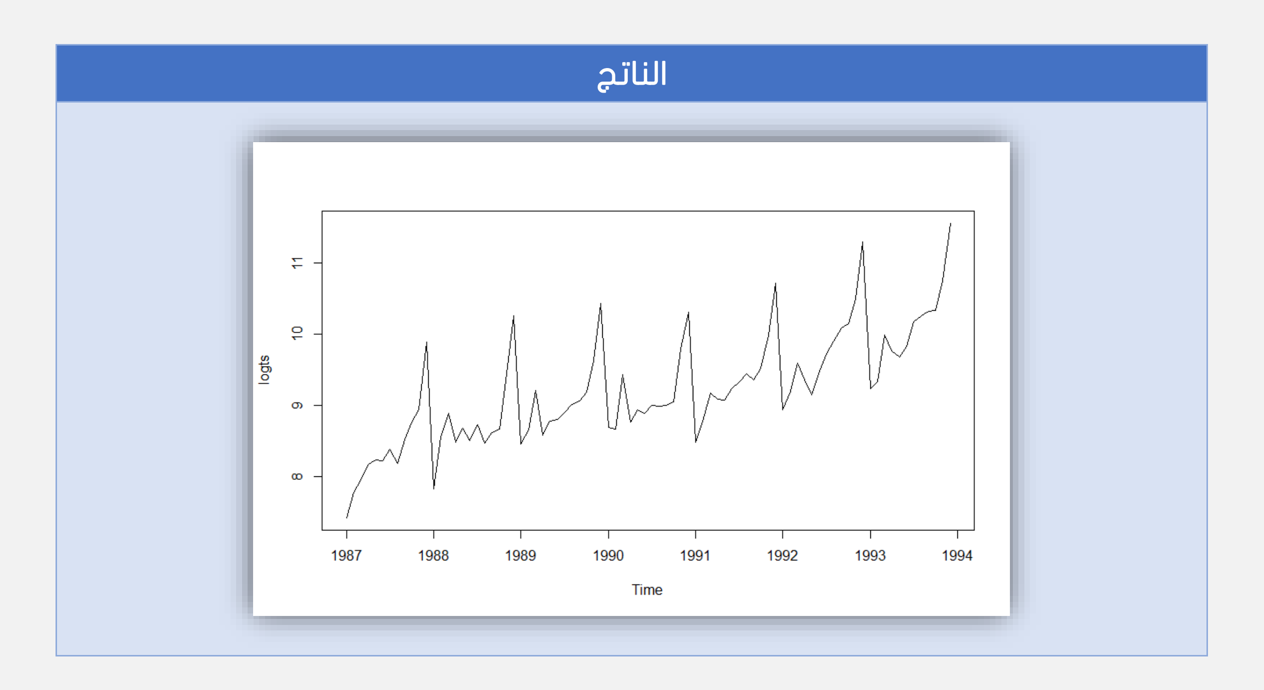
التعليمة
souvenir <- scan("http://robjhyndman.com/tsdldata/data/fancy.dat") souvenirts <- ts(souvenir, frequency=12, start=c(1987,1)) plot.ts(souvenirts)
الناتج



نلاحظ أن قيمة السلسلة الزمنية كبيرة جداً، مما يدلنا أن السلسلة الزمنية مكونة من مركباتها بواسطة الجداء، وبالتالي يجب ردها إلى سلسلة جمع بأخذ اللوغاريتم كما يلي:

التعليمة
logts<-log(souvenirts) plot.ts(logts)

الناتج



والآن سنشكل نموذج هولت ووينترز بالشكل:

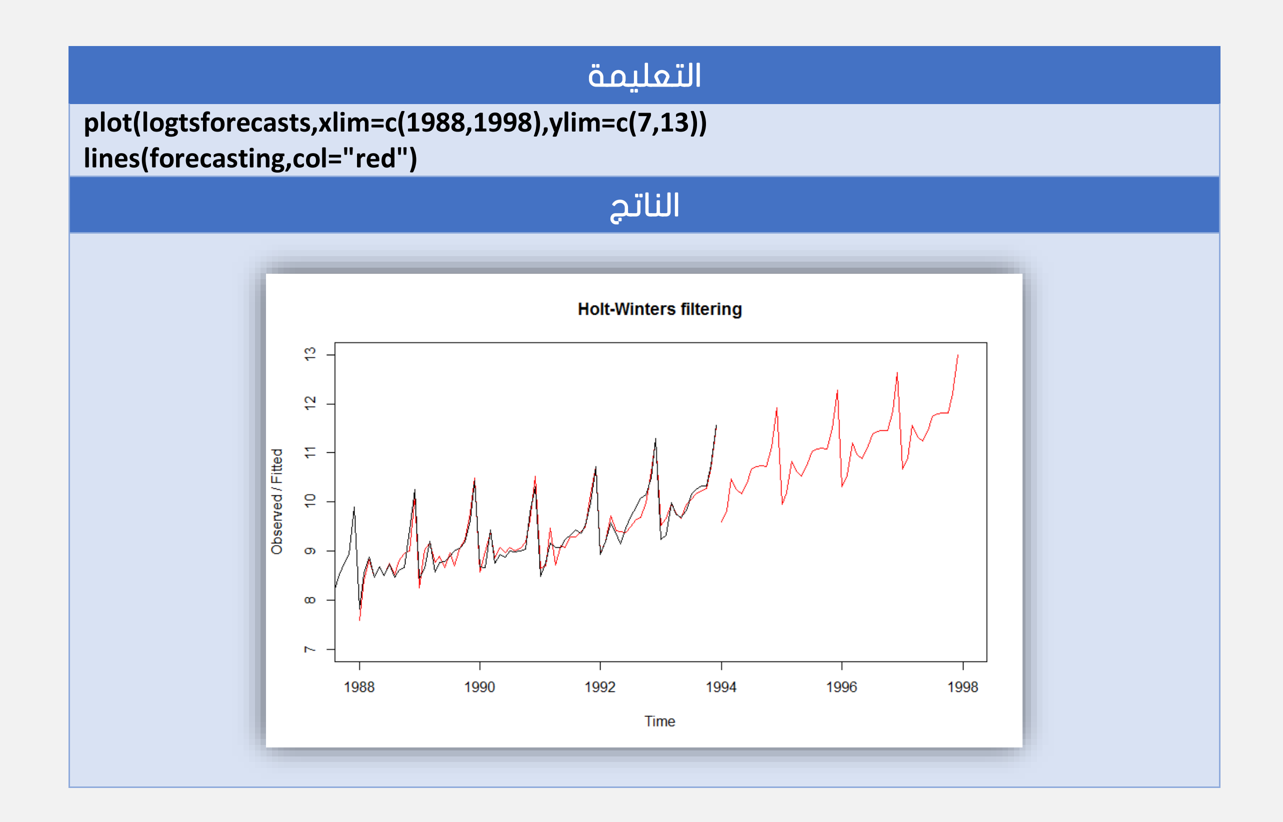
التعليمة

```
logtsforecasts<-HoltWinters(logts)
logtsforecasts
```

الناتج

```
Holt-Winters exponential smoothing with trend and additive seasonal component.
Call:
HoltWinters(x = logts)
Smoothing parameters:
 alpha: 0.413418
 beta : 0
 gamma: 0.9561275
Coefficients:
          [,1]
a   10.37661961
b    0.02996319
s1  -0.80952063
s2  -0.60576477
s3   0.01103238
s4  -0.24160551
s5  -0.35933517
s6  -0.18076683
s7   0.07788605
s8   0.10147055
s9   0.09649353
s10  0.05197826
s11  0.41793637
s12  1.18088423
```

يمكننا أيضاً رسم السلسلة المتنبأة والقيام بالتنبؤ بنفس الأسلوب السابق، حيث قمنا بالتنبؤ بقيم السلسلة الزمنية لـ 48 شهراً لاحقاً، أي 3 سنوات، وبالتالي سنكتب:



التعليمة
plot(logtsforecasts,xlim=c(1988,1998),ylim=c(7,13)) lines(forecasting,col="red")
الناتج

يجب أيضاً التحقق من عدم وجود ارتباط ذاتي بين الرواسب، ويجب التحقق من توزيع الرواسب طبيعياً بمتوسط صفري، وسنترك هذه الأمور للقارئ.

نماذج الانحدار الذاتي والمتوسطات المتحركة التكاملية لبوكس وجينكينز (Box – Jenkins Autoregressive Integrated Moving Averages Models):

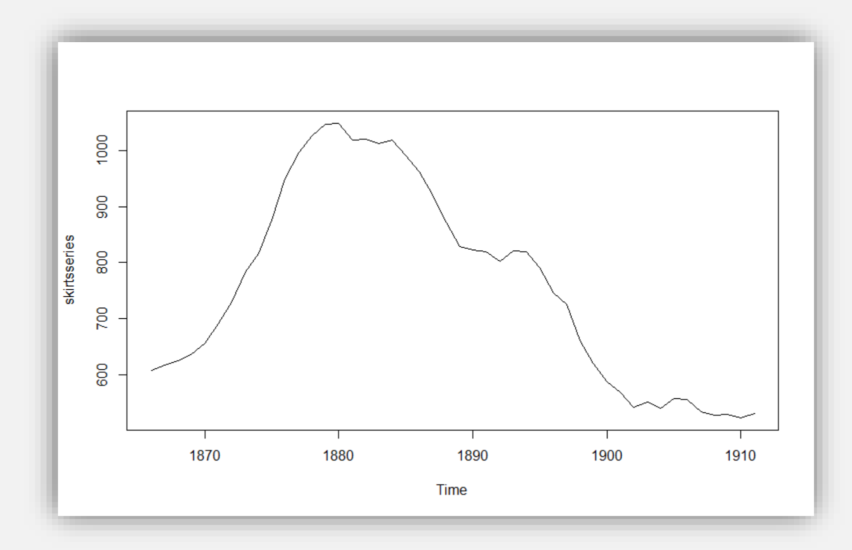
نماذج التمهيد الأسي تشترط عدم وجود ارتباط ذاتي بين الرواسب كما تشترط أن للرواسب توزيعاً طبيعياً بمتوسط صفري وتباين ثابت، لكنها لا تشترط أي شيء حول وجود ارتباط ذاتي بين قيم السلسلة الزمنية بحد ذاتها، وإن أخذ هذا الارتباط بعين الاعتبار يحسن كثيراً من دقة التنبؤ في كثير من السلاسل الزمنية.

تشترط سلاسل ARIMA أن تكون السلسلة الزمنية مستقرة، وإذا كانت السلسلة ليست مستقرة فيمكن تحويلها إلى سلسلة مستقرة بأخر تحويلة الفروق للسلسلة الزمنية، وهذا ما سنبدأ به:

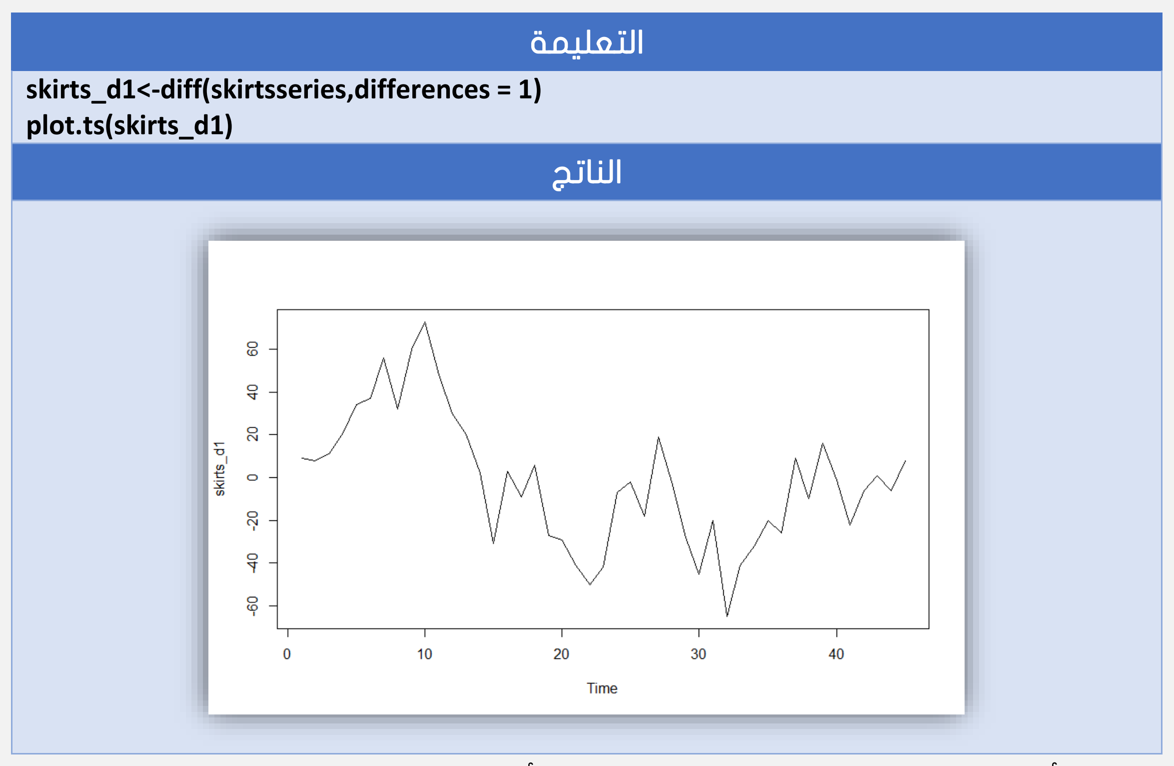
تحويلة الفرق للسلسلة الزمنية (Differencing of Time Series):

نرمز لنموذج ARIMA بشكل عام بالرمز ARIMA(p,d,q)، المركبة d تعني مرتبة الفروق التي تم إجراؤها على السلسلة الزمنية حتى أصبحت مستقرة، فلو أجرينا فرقاً واحداً لكتبنا ARIMA(p,1,q)، وهكذا...، ويمكن أخذ الفرق للسلسلة الزمنية باستخدام التابع diff().

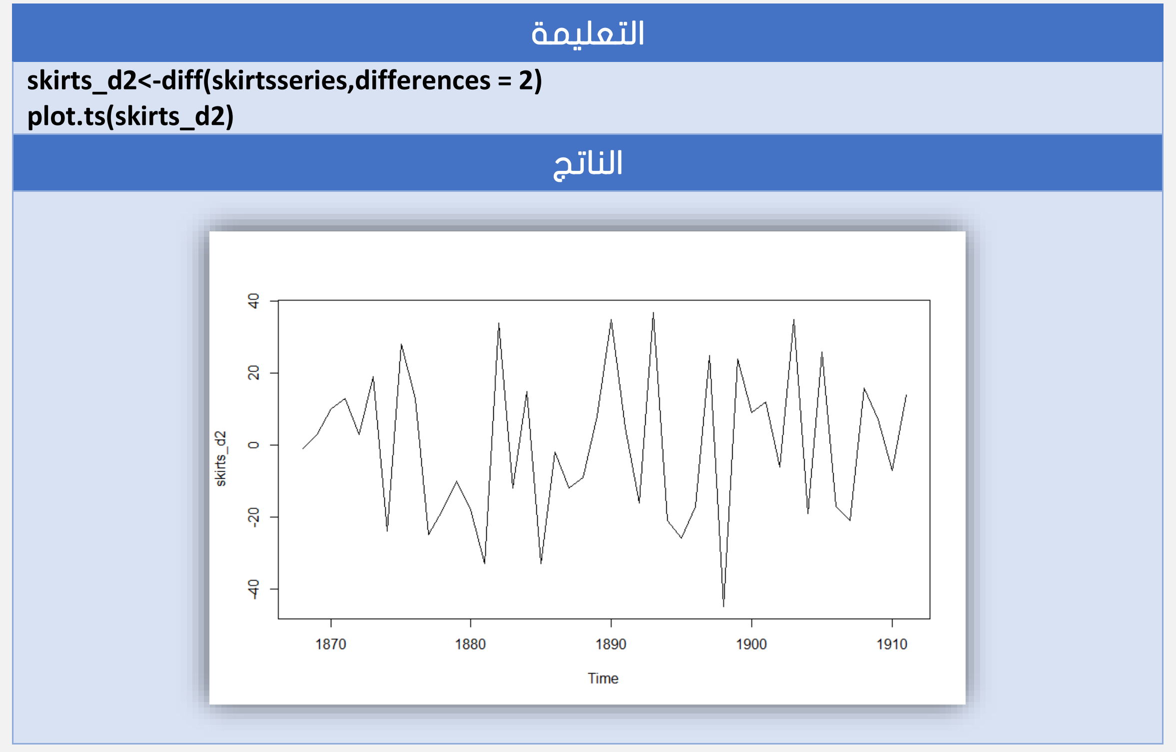
لو نظرنا إلى سلسلة قياس الخصر للتنانير التي استعرضناها سابقاً لوجدنا أنها ليست مستقرة:



بأخذ الفرق من المرتبة الأولى نجد:

التعليمة
skirts_d1<-diff(skirtsseries,differences = 1) plot.ts(skirts_d1)
الناتج


ونلاحظ أن السلسلة لا زالت غير مستقرة لذلك **س**نأخذ الفرق من المرتبة الثانية:

التعليمة
skirts_d2<-diff(skirtsseries,differences = 2) plot.ts(skirts_d2)
الناتج


ونلاحظ أن السلسلة أصبحت مستقرة.

الخطوة الآتية هي معرفة قيم كل من p,q ومعناهما في ARIMA(p,d,q):

اختيار نموذج ARIMA الملائم (Selecting Appropriate ARIMA Model):

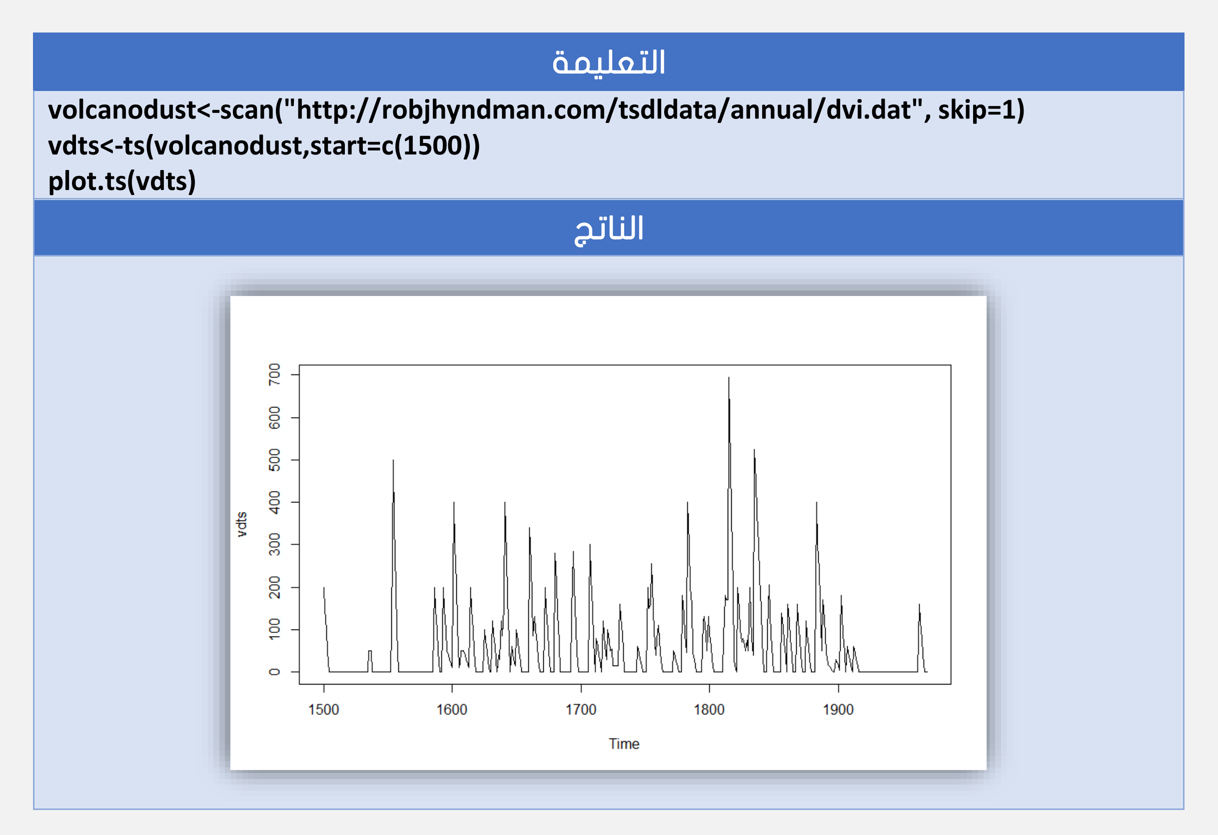
اختيار نموذج ARIMA الملائم يعني اختيار قيم كل من p,q المناسبتين، وذلك يتم بالاعتماد على كل من دالة الارتباط الذاتي ودالة الارتباط الذاتي الجزئي باستخدام التابعين acf() و pacf().

-نستخدم النموذج AR(p)=ARIMA(p,0,0) عندما تكون قيم دالة الارتباط الذاتي الجزئي تساوي الصفر بعد الفجوة رقم p، وتقترب قيم دالة الارتباط الذاتي من الصفر تدريجياً.

-نستخدم النموذج MA(q)=ARIMA(0,0,q) عندما تكون قيم دالة الارتباط الذاتي تساوي الصفر بعد الفجوة رقم q، وتقترب قيم دالة الارتباط الذاتي الجزئي من الصفر تدريجياً.

-نستخدم النموذج ARMA(p,q)=ARIMA(p,d,q) إذا كانت قيم كل من دال الارتباط الذاتي ودالة الارتباط الذاتي الجزئي تقتربان من الصفر تدريجياً.

لنأخذ السلسلة الزمنية الآتية التي تمثل الغبارات البركانية في نصف الكرة الأرضية الشمالي من العام 1500 وحتى 1969: http://robjhyndman.com/tsdldata/annual/dvi.dat



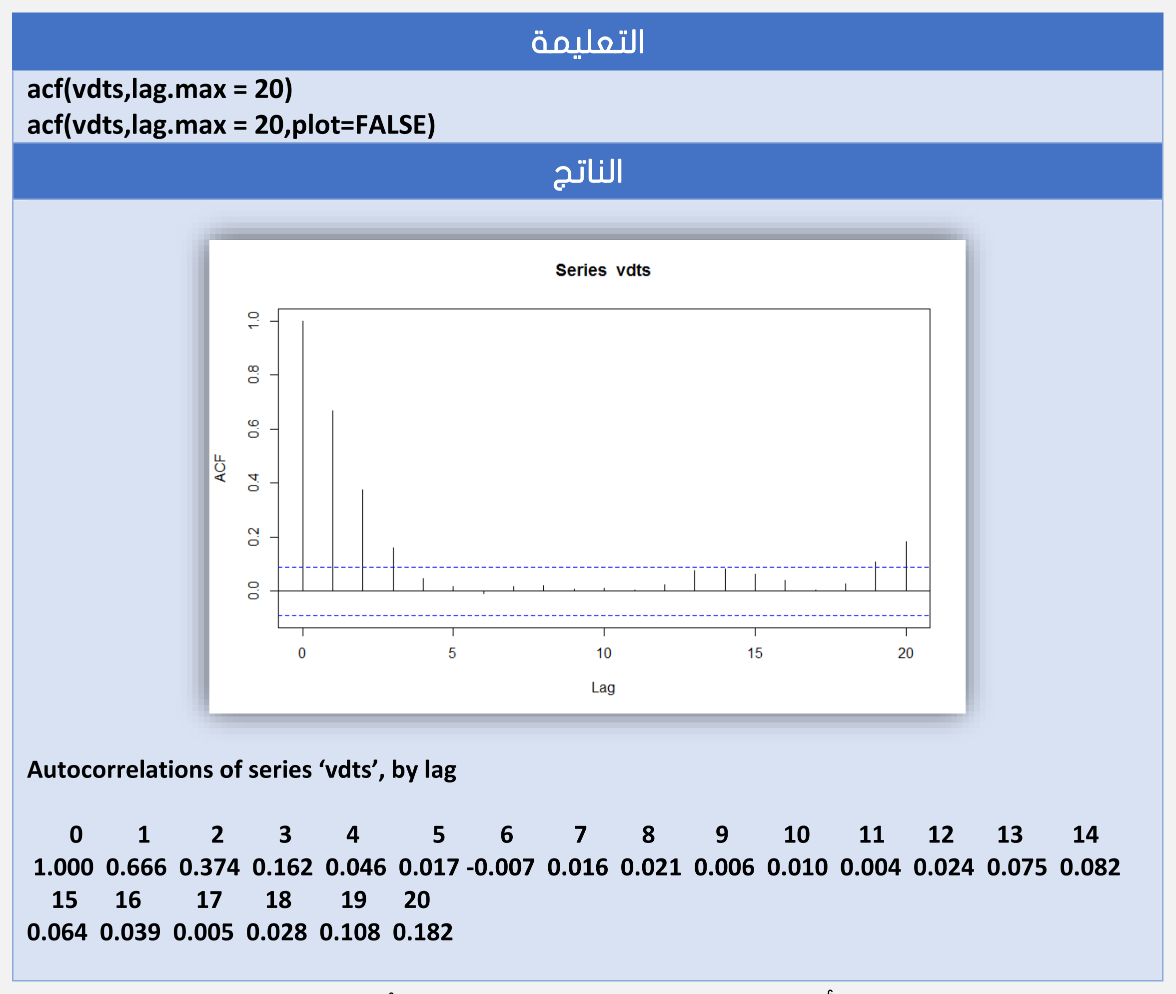
التعليمة

```
volcanodust<-scan("http://robjhyndman.com/tsdldata/annual/dvi.dat", skip=1)
vdts<-ts(volcanodust,start=c(1500))
plot.ts(vdts)
```

الناتج

نلاحظ أن السلسلة الزمنية السابقة مستقرة في المتوسط والتباين حيث يبين أن مستواها وتذبذبها بشكل عام ثابت على طول الزمن.

سنقوم الآن برسم تابع الارتباط الذاتي للسلسلة السابقة وإيجاد قيم الارتباطات الذاتية:



التعليمة

```
acf(vdts,lag.max = 20)
acf(vdts,lag.max = 20,plot=FALSE)
```

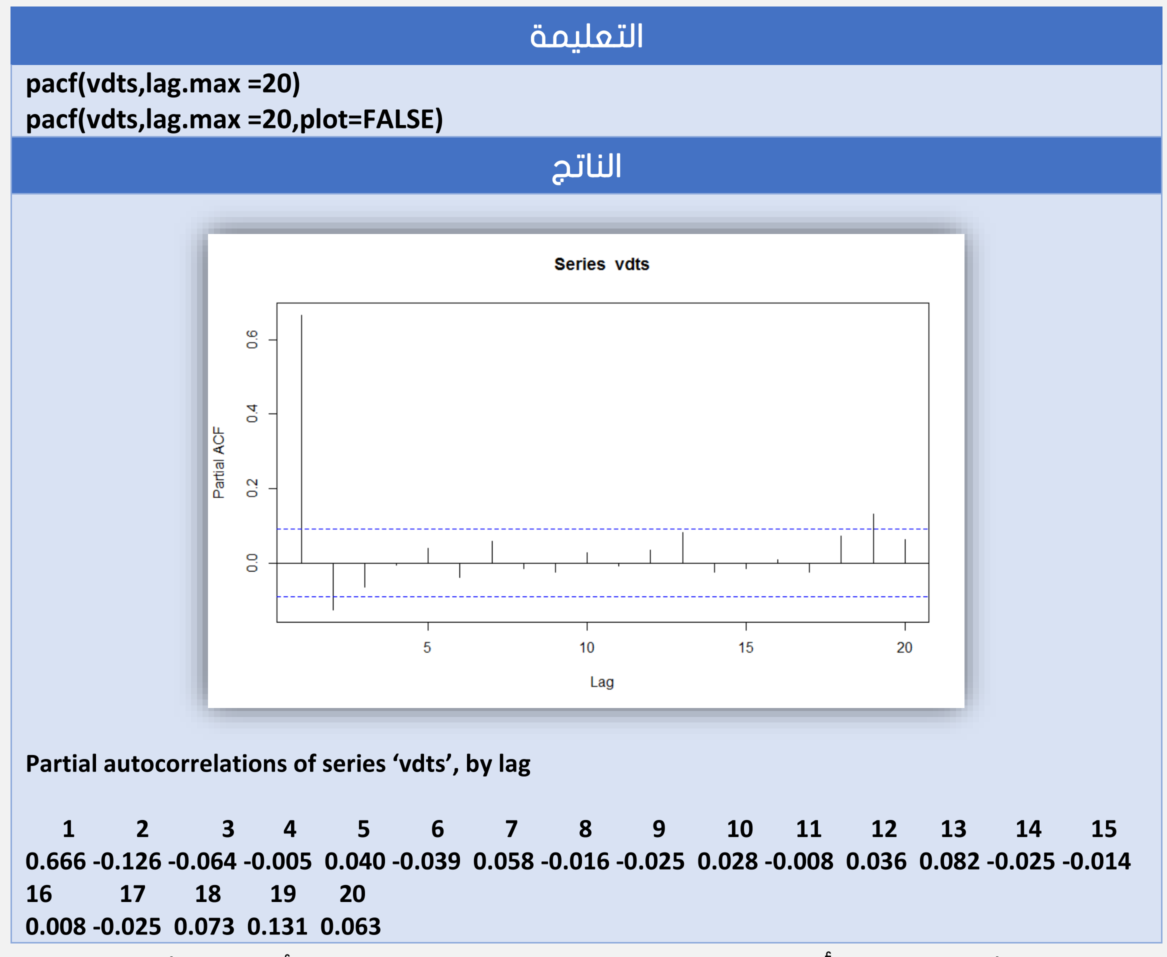
الناتج

```
Autocorrelations of series 'vdts', by lag

    0     1     2     3     4     5      6     7     8     9    10    11    12    13    14
1.000 0.666 0.374 0.162 0.046 0.017 -0.007 0.016 0.021 0.006 0.010 0.004 0.024 0.075 0.082
   15    16    17    18    19    20
0.064 0.039 0.005 0.028 0.108 0.182
```

نلاحظ من الشكل السابق أن قيم الارتباط الذاتي للفجوات الأولى والثانية والثالثة تجاوزت خط المعنوية المسموح به ثم بدأت تنعدم من الفجوة الرابعة وما بعدها.

سنقوم الآن برسم تابع الارتباط الذاتي الجزئي للسلسلة الزمنية السابقة وإيجاد قيم الارتباطات الذاتية الجزئية لها:



التعليمة
pacf(vdts,lag.max =20) pacf(vdts,lag.max =20,plot=FALSE)
الناتج

Partial autocorrelations of series 'vdts', by lag

```
     1      2      3      4      5      6      7      8      9     10     11     12     13     14     15 
 0.666 -0.126 -0.064 -0.005  0.040 -0.039  0.058 -0.016 -0.025  0.028 -0.008  0.036  0.082 -0.025 -0.014 
    16     17     18     19     20 
 0.008 -0.025  0.073  0.131  0.063 
```

نلاحظ من الشكل السابق أن قيم الارتباط الذاتي الجزئي للفجوات الأولى والثانية تجاوزت خط المعنوية المسموح به ثم بدأت تنعدم من الفجوة الثالثة وما بعدها.

نحن الآن في حيرة أمام ثلاثة نماذج ممكنة:

ARIMA(2,0,0): لأن دالة الارتباط الذاتي الجزئي انعدمت اعتباراً من الفجوة الثالثة.

ARIMA(0,0,3): لأن دالة الارتباط الذاتي انعدمت اعتباراً من الفجوة الرابعة.

ARIMA(p,0,q): كون كل من دالتي الارتباط الذاتي والارتباط الذاتي الجزئي تنعدمان تدريجياً.

الحل هو استخدام الدالة ()auto.arima حيث تعطي هذه الدالة أفضل نموذج ARIMA لملاءمة بيانات السلسلة الزمنية:

التعليمة

```
myModel<-auto.arima(vdts)
myModel
```

الناتج

```
Series: vdts
ARIMA(1,0,2) with non-zero mean

Coefficients:
         ar1     ma1     ma2    mean
      0.4723  0.2694  0.1279  57.5178
s.e.  0.0936  0.0969  0.0752   8.4883

sigma^2 estimated as 4897:  log likelihood=-2661.84
AIC=5333.68   AICc=5333.81   BIC=5354.45
```

تبين التعليمة السابقة أن أفضل نموذج هو ARIMA(1,0,2) وتعطينا معالمه والأخطاء المعيارية في المعالم، كما توفر لنا أيضاً معايير كل من AIC و AICc و BIC لمقارنة عدة نماذج واختيار النموذج الأفضل.

يمكننا التنبؤ بالقيم المستقبلية للسلسلة الزمنية بنفس الطريقة المتبعة سابقاً، أو بالطريقة الآتية (فرضاً سنتنبأ بعشر قيم مستقبلية):

التعليمة

```
forecast(myModel,h = 10)
```

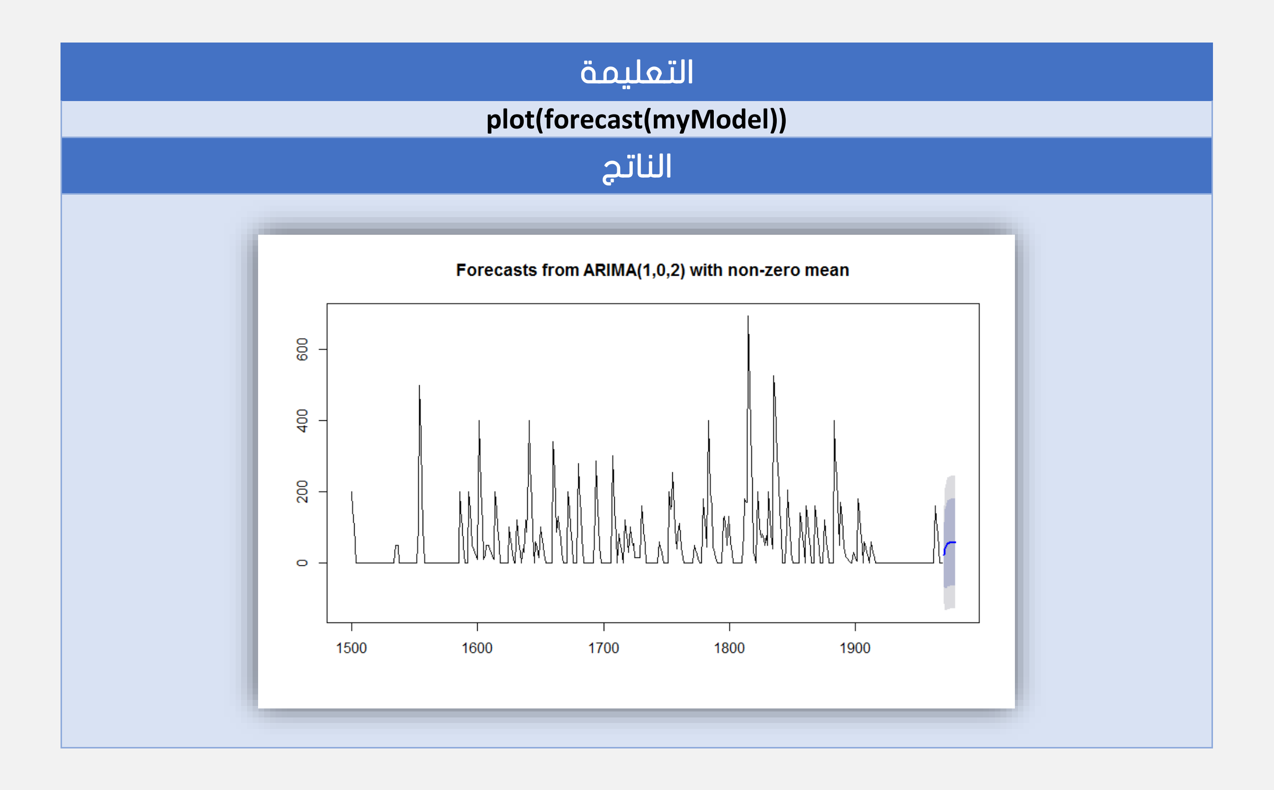
الناتج

Point	Forecast	Lo 80	Hi 80	Lo 95	Hi 95
1970	22.67720	-67.00065	112.3550	-114.4732	159.8276
1971	38.42748	-73.22746	150.0824	-132.3340	209.1890
1972	48.50091	-71.10909	168.1109	-134.4268	231.4286
1973	53.25886	-68.05471	174.5724	-132.2742	238.7920
1974	55.50617	-66.18421	177.1965	-130.6032	241.6155
1975	56.56763	-65.20665	178.3419	-129.6701	242.8053
1976	57.06898	-64.72400	178.8620	-129.1973	243.3353
1977	57.30579	-64.49137	179.1029	-128.9669	243.5785
1978	57.41764	-64.38045	179.2157	-128.8565	243.6917
1979	57.47046	-64.32783	179.2688	-128.8040	243.7449

نلاحظ أن R قدم لنا تنبؤات لعشر سنوات مع مجالات تنبؤ بمستوى 80% و 95%.

ماذا لو أردنا رسم النموذج مع القيم المتنبأة المستقبلية؟

ذلك يتم بالشكل الآتي:



التعليمة
plot(forecast(myModel))
الناتج

أخيراً نقول إن السلاسل الزمنية علم قائم بحد ذاته، وكذلك كل فصل في هذا الكتاب هو علم قائم بحد ذاته، ولن نستطيع أن نعطي هذه العلوم حقها في كتاب متواضع كهذا، لكن الهدف هو وضعك عزيزي القارئ في بداية الطريق لتكمل أنت إبحارك كما تحتاج وتشاء.

---------------------------------- انتهى -------------------------------------

المراجع

1 A Beginner's Guide to R, 2009
2 A first Course in Statistical Programming with R, 2007
3 A Handbook of Statistical Analyses Using R, 2006
4 A Little Book of R for Time Series, Avril Coghlan, 2017
5 Advanced R, Hadley Wickham, 2014
6 An R Companion to Applied Regression, 2010
7 Applied Econometrics with R, 2008
8 Applied Meta-Analysis with R, 2013
9 Bayesian Computation with R, Jim Albert, 2007
10 Beginning R: An Introduction to Statistical Programming, Larry Pace, 2012
11 Data Analysis and Graphics: Using R - an Example-Based Approach. Cambridge Series in Statistical and Probabilistic Mathmatics, 2003
12 Data Analysis with R, Tony fischetti, 2015
13 Data manipulation with R, Phil Spector, 2008
14 Data Mining with R: Learning with Case Studies, Luis Torgo, 2010
15 Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery, Graham Williams, 2011
16 Discovering Statistics Using SPSS, Andy field, 2000
17 Efficient R Programming: A Practical Guide to Smarter Programming, 2016
18 Extending R, John Chambers, 2016
19 Extending the Linear Model with R, Julian J. faraway, 2004
20 Graphical Models with R, 2012
21 Hands-On Programming with R: Write Your Own functions and Simulations, Garrett Grolemund, 2014
22 Introducing Monte Carlo Methods with R, 2009
23 Introductory statistics with R, Peter Dalgaard, 2002
24 Introductory Time Series with R, 2009

25 Learning R: A Step-by-Step function Guide to Data Analysis, Richard Cotton, 2013

26 Learning RStudio for R Statistical Computing, 2012

27 Mastering Predictive Analytics with R, Rui Miguel forte, 2015

28 Practical Data Science with R, Nina Zumel, 2014

29 R Cookbook, Paul Teetor, 2011

30 R for Data Science, 2016

31 R for Dummies, 2012

32 R for Everyone: Advanced Analytics and Graphics, Jared P. Lander, 2013

33 R Graphics Cookbook, Winston Chang, 2012

34 R Graphics, Paul Murrell, 2005

35 R in a Nutshell, Joseph Adler, 2009

36 R in Action: Data Analysis and Graphics with R, Robert Kabacoff, 2011

37 Reproducible Research with R and R Studio, Second Edition, Christopher Gandrud, 2013

38 Software for Data Analysis: Programming with R, John Chambers, 2008

39 Statistical Analysis with R, John M. Quick, 2010

40 Statistical Computing with R, Maria L. Rizzo, 2007

41 Statistics: An Introduction Using R, Michael J Crawley, 2005

42 The Art of R Programming: A Tour of Statistical Software Design, Norman Matloff, 2011

43 The R Book, Michael J Crawley, 2007

44 The R Inferno, Patrick Burns, 2012

45 Time Series Analysis with Applications in R, Jonathan D. Cryer, Kung-Sik Chan, 2008

46 Using R for Introductory Statistics, John Verzani, 2004

Statistical Programming Language



by

Mohamed Bisher Zeina

1st Edition